RL-TR-96-14
In-House Report
March 1996

THE IMPACT OF A MACHINE-READABLE LEXICON ON A PRINCIPLE BASED PARSER

Michael L. McHale

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Rome Laboratory
Air Force Materiel Command
Rome, New York

This report has been reviewed by the Rome Laboratory Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RL-TR-96-14 has been reviewed and is approved for publication.

APPROVED: Northrup Fowler

NORTHRUP FOWLER III
Chief, Software Technology Division
Command, Control & Communications Directorate

FOR THE COMMANDER: John A. Graniero

JOHN A. GRANIERO
Chief Scientist
Command, Control & Communications Directorate

If your address has changed or if you wish to be removed from the Rome Laboratory mailing list, or if the addressee is no longer employed by your organization, please notify Rome Laboratory/C3CA, Rome, NY 13441. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1996	In-House

4. TITLE AND SUBTITLE
THE IMPACT OF A MACHINE-READABLE LEXICON ON A PRINCIPLE BASED PARSER

5. FUNDING NUMBERS
PE - 62702F
PR - 5581
TA - 27
WU - TK

6. AUTHOR(S)
Michael L. McHale

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Rome Laboratory/C3CA
525 Brooks Rd.
Rome, NY 13441-4505

8. PERFORMING ORGANIZATION REPORT NUMBER
RL-TR-96-14

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Rome Laboratory/C3CA
525 Brooks Rd.
Rome, NY 13441-4505

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Rome Laboratory Project Engineer: Michael L. McHale, C3CA, (315)330-1458

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The goal of this research was to provide the groundwork for an investigation of an approach to domain independent semantic processing; the combination of a principle based parser (PBP) with a semantically enhanced machine-readable dictionary (MRD). The parser is an implementation of Chomsky's Government-Binding theory and therefore provides complete syntactic coverage. The coverage of a parsing system is, however, ultimately a function of the size and richness of its lexicon. To provide both size and richness, the lexicon for the system was extracted from Longman's Dictionary of Comtemporary English and semantically enhanced using Roget's International Thesaurus. Increased lexical richness increases system coverage but it may decrease the efficiency of the parser. Therefore, this research investigated the impact of using an MRD as the lexicon for a PBP.
The results show that an MRD can indeed be used with a PBP though the larger, more ambiguous lexicon requires controls in the parser to avoid producing a large forest of candidate parse trees. With such controls, the impact of the larger lexicon becomes no greater for a PBP than for a traditional phrase structure grammar (ex., ATN, APSG) dealing with lexical ambiguity.

14. SUBJECT TERMS
Natural Language Processing, computational Lexicography

15. NUMBER OF PAGES
110

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	U/L

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

Contents

Figures and Tables

Chapter 1
Introduction

Language is a perpetual Orphic song,
Which rules with Daedal harmony a throng of thoughts and forms,
which else senseless and shapeless were.

P.B. Shelley

This research is concerned with a computational linguistic method of processing text with the ultimate goal of improving the capabilities of information retrieval in providing relevant information and texts to users. The ability to index natural language text, using the concepts and connecting relations that occur in these texts, would result in an improvement in the quality of information over the simple word occurrence approaches currently in use. This premise was first championed years ago (ex., Sparck Jones 1974, Braun and Schwind 1976). Since then, parts of the premise have been demonstrated by a variety of researchers (ex., Dillon 1983, Wang et al. 1985, Croft and Lewis 1987). To fully demonstrate the premise though, a natural language processing (NLP) system with the ability to classify and categorize concepts and relations independently of domains[1] is required. That is, the domain independent nature of general information retrieval requires that a full demonstration of the premise include domain independent semantic processing. This type of semantic processing is well beyond the current capabilities of NLP. What *is* currently possible is domain independent syntactic processing, with semantic processing possible only in limited domains. This is insufficient for general information retrieval purposes. Those purposes could be better served if domain independent semantics could be added to a broad coverage syntactic system.[2] The immediate goal of the current research is the investigation of the first step to adding that semantic component.

---

[1] The word *domain* is being used throughout this research to mean *a field of study or interest*. It can be construed as a field in general (ex., economics, law, medicine) or as sub-fields (ex., obstetrics, corporate take-overs). It is not being used in its technical sense as is common in linguistics and mathematics.

[2] It is not necessary to use full syntactic analysis to do NLP for IR. The DR-LINK system (Liddy and Paik 1992, Myaeng et al. 1994) for instance, uses a more restrictive analysis of syntax (clause bracketing) along with lexical, semantic and discourse knowledge. The current research, though, is concerned strictly with full syntactic analysis.

Importance of the Problem

The connection between information retrieval (IR) and NLP seems quite intuitive. It appears fairly obvious that if a system is to determine and provide relevant information to a user the system should understand what that information is and therefore it is only natural that NLP would be very useful for IR systems. Over the lifetime of NLP a number of researchers have tried using NLP techniques to enhance the retrieval process.[3]

A major problem that surfaces for these systems is that information retrieval works with large heterogeneous data bases of unconstrained text whose very nature prohibits the use of domain dependent NLP techniques. Domain independent syntactic processing is possible, but as numerous studies have shown (ex., Fagan 1988, Smeaton 1989) "syntax without semantics isn't powerful enough" for IR (Salton 1991). Thus most successful NLP-based IR research has been limited to small domains where semantic information can be exploited.

The importance of IR will continue to grow as more and more data become available to more and more users (ex., in CD-ROMs for home computers and on the Internet and World Wide Web). The added demands brought by these untrained yet information hungry users are ill met by current IR practices of requiring the user to provide terms or keywords that hopefully occur in most of the relevant documents (*recall*) and only those documents (*precision*). An IR system that expects unsophisticated users to combine these keywords into an effective Boolean search string (ex., A **and** B **or** C **not** D) is doomed to failure. As Willem Scholten, the Director of Computer Systems and Research at Columbia University stated (Scholten 1993),

---

[3] Chapter Three will cover some of these systems. See also (Sparck-Jones and Kay 1973), (Becker 1981) and (Fagan 1988) for further reviews of the literature.

"Free-text searches using natural language queries are more useful and accurate than keyword searches. A free-text system does not require users to have prior knowledge of the discipline they are searching and allows users to search the full text of documents for arbitrary combinations of words."

NLP techniques thus allow naïve users a more "intuitive" means to communicate with IR systems. To be natural and intuitive, though, these NLP systems will require domain independent semantics with the ability to search using derived conceptual meanings and not just the keywords provided by the user.

While the challenge of naïve users is becoming more pervasive, of greater importance are the IR applications where "good enough" is insufficient. These are the applications that contain vital information and the retrieval of all the relevant information may be a matter of life or death. Systems that are involved with national defense, medicine and law, for example, cannot be satisfied with retrieving **most** of the relevant information but must strive to retrieve **all** the information germane to the user's problem. At the same time it is important that these systems do not inundate the user with irrelevant information as very few users have the time or will make the effort to wade through much information that is worthless to them. The only way that IR will be able to meet this demand will be by advancing the state-of-the-art in both conventional and novel retrieval techniques. In the near future this probably will mean using conventional IR techniques with high recall combined with semantic processing to weed out those sources of information likely to be non-relevant. As NLP semantic processing becomes more robust it should eventually be possible to index and retrieve information based on the ideas in the text and not just on the occurrence of any given word or phrase. This will facilitate the retrieval of information from disparate sources. For example, information relevant to NLP comes from the fields of psychology, philosophy, linguistics, mathematics and computer science. Each of these fields has its own vocabulary and thus searching for information using current techniques requires the rephrasing of

the same query into each of these different vocabularies[4] or sublanguages when searching their databases. The ability to retrieve on the ideas in the documents would not place this rephrasing burden on the user. Improvements in NLP technology will significantly affect IR in these and other ways, but advances in NLP technology will also have an impact on a large number of other fields.

Other Users

IR is not the only potential consumer of advanced NLP techniques. There are a number of other obvious users.

The creation of truly intelligent interfaces will require the processing of natural language. The understanding of written text may well play a limited role in these interfaces as few users would care to interact with a machine by typing long, complete sentences. However, the role for NLP will be pervasive both for speech understanding and for text and speech generation. A system that produces and understands speech will be able to interact with users on a much more human level than point and drag systems do. For speech systems to rise above the gimmick level, however, they must have an extensive NLP component including both semantics and pragmatics.

Intelligent interfaces that use NLP must also be capable of working in a variety of languages. To maximize utility, intelligent systems should be multi-lingual and translate freely among languages. This capability would provide access to these systems to a much wider audience. To be "fluent," machine translation systems have to deal with semantic issues of language. Current systems work well in restricted domains and with highly stylized texts (ex., translation of technical manuals, weather reports). However, to make these systems usable across domains

---

[4] For example, using the term *NLP* works fine for linguistics and computer science but returns information on Neuro-Linguistic Programming in psychology texts.

will require domain independent semantics. It is an approach to this type of semantics that is the focus of the current research.

Other potential uses include intelligent tutoring, automatic abstracting, command and control systems, database query, expert systems, management information systems, and information dominance. These applications would also benefit from any advances in the ability to handle lexical semantics independently from the domain.

Goal and Approach

The ultimate desire of IR researchers using NLP is a system that "understands" text to the point where it can extract the ideas, not just keywords, from the text. A logical first step to obtaining that goal is to expand NLP's capabilities in semantic processing. To that end, the long-range goal of our research is to investigate a methodology for adding lexical semantics to a broad coverage syntactic system. The way that this is envisioned is straightforward. A principle based parser (Berwick and Fong 1990) is being coupled with a machine-readable dictionary (MRD). This combination results in a domain independent syntactic system with good lexical coverage. A simple semantics will then be provided by extracting thematic relations (θ-roles) for the verbs using the Case Grammar[5] Matrix Model (Cook 1989). The result of this extraction process will be a θ-role frame for each verb sense. These frames are simply a pattern of the roles associated with each verb. For instance, for intransitive verbs the frame would consist of *verb[role1]*, and for transitive verbs it would be *verb[role1, role2]*, where *role1* represents the role of the subject and *role2*

---

[5] Case Grammar (Fillmore 1968, 1977) is a simple and elegant approach to semantics. This formalism identifies "who does what to whom" in a sentence. For instance, *Pat* is the AGENT in the sentence, "Pat ate lunch", so Pat is said to have the Case Role of AGENT. Throughout this research, these roles will be referred to as θ-roles.

that of the object. These roles are not inherently part of the noun. For instance, a person may be the AGENT, BENEFACTOR, or EXPERIENCER in a sentence. The verbs determine the roles of their objects. Therefore, only the verbs need to be analyzed. After they are extracted, the θ-role frames can be enhanced by adding to the frame a pointer to the verb's location in a general, conceptual hierarchy. This provides a richer lexical semantics that should prove useful to IR and other applications that require some domain independent semantic processing.

The premise behind this approach is that thematic relations (ex., AGENT, THEME,[6] EXPERIENCER) can form a solid link between syntax and semantics. The premise will be tested by investigating the ability of thematic relations to link a principle based parser (based on Government-Binding theory) and lexical semantics. Principle based parsing was chosen for this research because it is based on language independent (and therefore domain independent) principles and guarantees theoretical syntactic coverage of a language. Furthermore, one of the principles used governs the proper assignment of thematic relations and thus principle based parsing is a natural choice for this research. Principle based parsing and Government-Binding theory will be covered in more detail in chapter two.

Semantics

The choice of lexical semantics is almost as natural. There are at least three viewpoints on semantics. The first views words as filling some function in a sentence and therefore views semantics as the logical delineation of those functions. This is the viewpoint taken by Montague (Montague 1972, 1974, Dowty et al. 1981)

---

[6] THEME has been referred to as OBJECT or PATIENT by various researchers. THEME is being used here for two reasons: first, it follows that traditionally used in principle based parsing (Gruber 1965, Jackendoff 1990); second, the use of the term OBJECT is objectionable because in the current framework *object* is purely a syntactic structure. As Jackendoff has observed (1990:47) the names of the roles "are just convenient mnemonics for particularly prominent configurations."

and is known as *formal semantics*. In this formalism the semantics of a sentence is represented in first-order, modal, intensional or some other formal logic. The approach has been used successfully in various NLP systems, for example in the Chat-80 system (Pereira 1982). It is a truth-theoretic model, that is, sentences have a truth value associated with them that is a composition of the values of each word. The main drawback with this approach is that the function of a word is very often domain dependent and thus the resulting semantic system is also domain dependent. The dependence on domain can be seen as simply restricting the senses that a word can assume. For instance, *icing* in hockey infers the movement of an object and not something edible on cake. A different type of change in function occurs when the domain does not require certain classes of words (i.e., they become functionless). An example of this phenomenon is telegraphic or military message traffic in which prepositions, determiners and other, so-called, function words are deemed superfluous (ex., "ARRIVING SYRACUSE 4:00 AMERICAN FLIGHT 704"). Another problem with the formal semantics approach is that the truth-theoretic model inherent in the logical systems used in NLP does not work well with much of natural language. For instance the sentence "Pat ate the apple" can be seen as being either true or false. The related question, "Did Pat eat the apple?", however, has no truth value. A large number of speech acts are even further removed from the true-false dichotomy than simple questions are. Sarcasm, rhetorical questions, performative statements (ex., "I now pronounce you man and wife" where the statement causes a change in state and is not just true), indirect questions (ex., "Can you pass the salt?"), proverbs, greetings, and commands are all examples of common speech acts that do not easily conform to the truth-theoretic model. These difficulties, and the problem of scaling up to domain independence, preclude the use of formal semantics for this research.

Another viewpoint on semantics is that words have little or no meaning except in context. This viewpoint can be traced at least as far back as the *field semantics* of Trier (1932) and the writings in the 1930's of Wittgenstein (1934). It is also reminiscent of the *valeur* of de Sassure (1901). Computationally this approach has been tried in various forms including the Word Expert Parser (Rieger and Small 1968, Papegaiij 1986). This approach is not inherently domain dependent but the size of the lexicon needed for each entry effectively limits the domain for any given system. The amount of information needed as explicated in Rieger and Small (1968) greatly exceeds that which is found in MRDs so this approach to semantics was likewise not considered.

Thematic relations cut across these two viewpoints of semantics. Like formal semantics, θ-roles describe a function of the word in the sentence. Like field semantics, the relations work with the context of a sentence. It is a limited context because only the verb and its arguments are considered, though they must agree in type. For instance, a noun should be animate to accept the role of AGENT. This fusing of semantic approaches gives thematic relations greater flexibility than formal semantics. Yet the relations still deal very loosely with the meaning of the word. Since semantics is the *science of meanings*, or at least it was when Michel Bréal coined the word in 1883, a meaning oriented approach is warranted.

In this research, the thematic relations will be extended to include *lexical semantics* (Cruse 1986, Pustejovsky and Boguraev 1991). In the lexical semantics view, words do have an intrinsic meaning or meanings and the task is choosing among them. This parallels nicely the traditional dictionary view where the meaning of a word is defined, and the senses of a word are juxtaposed. A dictionary, then, provides a type of lexical semantics to users, with the selection of senses often aided by the inclusion of example sentences.

The meaning of a word is important, but for NLP systems the representation of the meaning is also an important consideration. The definitions could be retained as the representation, though doing so would require reprocessing the definition each time the meaning is needed by the system. This type of representation therefore seems less attractive to machine processing, where elimination of redundant processing increases speed, than it is for human users, where redundant processing increases comprehension. Instead of retaining the definitions, the use of antonyms or synonyms might be considered, though it is often difficult to find a precise term for each word (ex., what is the antonym of *kiwi fruit*? or the synonym of *assassinate*?). Another possible representation is by picking unambiguous labels for each sense of a word and placing these labels, and their corresponding words, into a categorization of conceptual space. This last approach sounds very abstract but is similar to what is commonly done in libraries with classification schemes (i.e., Dewey Decimal, Library of Congress). It is also the approach that was used by Peter Roget in the construction of his thesaurus (Roget 1852, fifth edition 1992).

In Roget's representation schema, which is used in our research, the meaning of a word is represented by its location in a conceptual hierarchy. Each location is unambiguously labeled with a category name (ex., *to struggle* is in the category *exertion*). The category names can be placed with the verb in the corresponding θ-role frame thus providing a pointer to the location in the conceptual hierarchy where the sense of the words fit. This pointer, in effect, provides lexical meaning directly to the frame. The approach uses the θ-role as a pivotal link between syntax and lexical semantics. It provides a precise, efficient representation for the NLP system while producing a meaningful abstraction for IR purposes. *Roget's International Thesaurus* will be covered in more detail in chapter two.

Research Question

The long-range goal of the current research is the investigation of a methodology for adding domain independent semantics to a broad coverage syntactic system. The main premise behind the method is that thematic relations can form a solid link between syntax and lexical semantics. Three components are required to test the premise: a broad coverage syntactic system consisting of a domain independent grammar and a large lexicon (MRD derived); thematic relations for the lexicon; and a form of lexical semantics for the lexicon. Consequently the research raises questions dealing with the interaction of the grammar and the lexicon; the ability to extract the thematic roles; and the extension of those roles to lexical semantics. Specifically the questions are: (1) How much impact does a large, general lexicon have on a principle based parser? (2) To what extent are θ-roles automatically derivable from a machine-readable dictionary? (3) To what extent can the θ-role frames be automatically placed into a conceptual hierarchy? The current research deals with just the first of these questions.

Underlying the foundation for these questions are two minor premises: (1) A suitable principle based parser can be written. (2) An MRD provides a good basis for an NLP lexicon. Since the validity of these statements is not self-evident, it is worthwhile to examine them in some detail.

The creation of a principle based parser is not a research issue as a dozen or so principle based parsers have already been written by a variety of researchers. Furthermore, as part of a feasibility study for this research, a prototype parser was written that included four of the major principles. That study evidenced the feasibility of writing a full parser using a hand-coded lexicon. So, in a very real sense, the first minor premise can be seen as true; the writing of a suitable parser is feasible. Extending a principle based parser to use anything but a hand-coded

lexicon remains a research issue[7] though, and is addressed in the first research question.

An initial issue one might raise when extending the parsing system to work with a new lexicon is, "Does the new lexicon provide sufficient information to meet the needs of the parser?" Since a principle based parser is in effect a lexical parser it requires a richer lexicon than some of the more common parsers (ex., augmented transition nets). The information that is required by a lexical parser includes, for instance, the syntactic category, lists of reflexive pronouns, and the marking of verbs with the number and types of arguments they require. Obviously some of this information is explicitly present in a general purpose dictionary along with information that the parser does not need (ex., pronunciation, etymology). Some of it, however, is only implicitly present, if at all. In previous research (Mc Hale 1991), the machine-readable version of *Longman's Dictionary of Contemporary English* (LDOCE 1987) was closely examined and was found to provide the vast majority of the information, both explicitly and implicitly, that is needed by the parser.[8] Thus the second minor premise is also true; LDOCE is a good first source of information for the lexicon.

A second issue regarding the replacement of the hand-coded lexicon with a more general one is particular to principle based parsers. Most of the parsers to date (ex., Abney and Cole 1985, Fong 1991, Correa 1988, Dorr 1987, Kuhns 1990) have been restricted to small domains with little lexical ambiguity. The introduction of lexical richness into this type of parser might lead to a virtual explosion of possible interactions among the principles and word senses. An explosion of this sort could

---

[7] While the PRINCIPAR parser (Lin 1994) uses an MRD it is not concerned with semantics. The parser also overly constrains the number of parse trees that it builds making it difficult to assess the impact of the lexicon.

[8] LDOCE was chosen for the present research based on a number of characteristics that make it especially attractive for an NLP lexicon. These characteristics are covered in more detail in chapter two.

have an impact on processing time, the number of parses produced, and generally degrade performance. The question of impact is answered by comparing processing characteristics of the parser using a single test corpus and varying only the lexicons. A carefully selected corpus (Appendix 1) was parsed first using a small, hand-coded lexicon. The corpus was chosen for its ability to test the principles used by the parser, and thus the syntactic coverage. This test not only ensures syntactic coverage but provides a baseline against which the parser can be tested when the switch to LDOCE is accomplished. The switch to LDOCE also allows the demonstration of the domain independent nature of the parser/LDOCE combination. The results for this question are given in Chapter Four.

The three questions address the main points raised by this approach of using thematic relations as the pivotal link between syntax and semantics. Thematic relations are an integral component of the syntactic parser and the use of that parser with a general lexicon is the subject of the current research. The second question addresses the extraction of the thematic relations from the lexicon. The resultant thematic relations are useful but provide an impoverished semantics at best. Indeed, this sort of subject, object, locative marking occurs not infrequently in the syntax of natural languages (for instance, in Japanese these roles are overtly syntactic in nature). The third question investigates a way to enhance the thematic relations to a much richer semantic representation than that required by either Government-Binding or Case Grammar theory. The result should prove useful for IR. The three questions taken together then provide one way of investigating an approach to domain independent semantic processing.

Contributions of the Research

The significance of this research is both practical and theoretical. The research should provide the basis of an NLP system that would be of practical interest to IR,

intelligent interfaces and a wide spectrum of other applications. Also of interest is the investigation of the control needed by a principle based parser when coupled with a large, ambiguous lexicon. Additionally, the extraction of θ-roles from lexical sources and the linking of those roles with the semantic hierarchy in Roget's are of interest. Of perhaps even greater theoretical significance is the question concerning the possibility of extending a principle based parser with this type of lexical semantics. While the semantic link has existed in the underlying Government-Binding theory for some time it has not been put to a serious test.

The intent of this research is to explore a principle based parser that uses a semantically enriched lexicon automatically derived from machine readable lexical sources. It is to be hoped that this would eventually lead to a domain independent, robust, NLP system suitable for IR and other information based uses. If the results help to illuminate such systems then there can be little doubt as to the success of the research.

Additionally, there are a number of intermediary results that would significantly contribute to contemporary research in NLP, principle based parsing and IR.

One benefit that will hopefully be a product of this research is a wider awareness of the computability of principle based parsing. This could be considered a benefit in a number of ways. There are several characteristics of the underlying Government-Binding theory that make it attractive as the basis for parsing. (a) The ability to create a "new" parser by simply setting a few parameters (ex., if adjectives generally come before or after the noun) that are germane to a given language or domain (ex., Dorr 1987). (b) The "grammar" is based on linguistic principles and is not simply a collection of rules – one rule for each syntactic structure encountered by the system designer. (c) The modularity of the principles allows for independent testing and maintenance of components, and (d) the size of the system is

considerably smaller than that of a rule based parser with comparable coverage. All of these features make a principle based parser attractive but the practical computability of principle based parsing has never been fully accepted. In theory it is at least as computable as some of the more popular rule-based approaches (Barton, Berwick and Ristad 1987). It is not clear, however, that a principle based parser faced with the richness of lexical ambiguity found in normal text would be as efficient as other systems. If this research can demonstrate a parser, using an MRD derived lexicon, running on domain independent text, then more researchers may use this formalism for future NLP applications. This increased usage could benefit NLP, and NLP applications, by decreasing the amount of time spent in creating full coverage parsers. It could also benefit linguistics by providing more computational testbeds for linguistic research.

Finally, this research anticipates contributing to the use of NLP techniques for IR research. If any of the techniques developed are in themselves useful, or if the research increases the awareness of the suitability of NLP for IR, then the research will be successful.

Overview

Chapter 2 has a discussion of the lexical resources and major components being used (*Longman's Dictionary of Contemporary English, Roget's International Thesaurus*, and Government-Binding theory) and thus provides background material for chapter 3. Chapter 3 covers some selected systems found in the literature in the areas of: NLP research for IR; principle based parsing; and the use of machine-readable lexical sources for NLP. Chapter 4 covers the methodology, testing and results used for measuring the impact of the use of an MRD with the parser. Chapter 5 provides a short summary and evaluation of the overall research with suggestions for further investigations.

Chapter 2

Background

Thirty spokes share the wheel's hub;
  It is the center hole that makes it useful.
Shape clay into a cup;
  It is the space within
  that makes it useful.
Cut doors and windows for a room;
  It is their emptiness
  that makes them useful.
Therefore profit from what is there;
  Utilize what is not.

Lao Tzu, *Tao Te Ching*

This chapter is divided into four main parts. The first part will give an overview of Government-Binding theory (the theoretical foundation for principle based parsing) and show why it is uniquely qualified as the basis for a wide coverage, context-independent grammar. The second part will discuss *Longman's Dictionary of Contemporary English* (LDOCE) which is being used as the basis for the lexicon of the principle based parser. Next the theory of thematic relations will be reviewed. These relations (θ-roles) are required by the parser but are only implicitly present in LDOCE. The chapter will conclude with an outline of *Roget's International Thesaurus* and its conceptual hierarchy, which is being used to enhance the semantics of the θ-roles to a point deemed usable by information retrieval.

Government-Binding Theory

It may be easiest to explain Government-Binding by comparing it to a traditional phrase structure grammar. A phrase structure grammar consists of a number of phrase structure rules, which provide alternative representations for a phrase. For instance, the phrase structure rule

sentence → noun-phrase verb-phrase

indicates that a (simple) sentence can be represented as a noun phrase followed by a verb phrase. A grammar is constructed by adding enough rules to allow complete sentences to be analyzed. For example, the grammar

sentence → noun-phrase verb-phrase

noun-phrase → noun

verb-phrase → intransitive-verb

would be sufficient for analyzing sentences such as *dogs bite* or *chickens sleep.* The analysis would produce a structure something like

sentence[noun-phrase[noun[dogs]] verb-phrase[intransitive-verb[bite]]].

To analyze the sentence *the dog bites* the rule for the noun phrase would have to be modified to allow determiners. This sort of modification is usually done by making the component, in this case determiners, optional. Optional components are denoted by parentheses.

noun-phrase → (determiner) noun

The modification could also be done simply by adding a separate rule. The analyst would then decide which rule to use for any given sentence. Repeated components are denoted by parentheses and asterisks. Thus,

noun-phrase → (determiner) (adjective)* noun

would allow any number of adjectives, such as *the big, red, steel door,* but not *the red, white and blue flag.* The latter phrase would take another rule or two to handle the conjunction. In general, phrase structure grammars work well for small domains. Larger domains generally have more variety in their language constructs and therefore have a larger grammar. Eventually the phrase structure grammar becomes so large as to become unwieldy. It is for these larger domains that Government-Binding theory becomes attractive.

Government-Binding (Chomsky 1981, 1982, 1986) is a principle-based syntactic theory that has been continuously developed and refined by repeated applications to a wide variety of languages (Berwick and Fong 1990).[1] These range from traditional NLP languages such as English and French, to very unrelated languages such as German, Portuguese, Japanese and Warlpiri (a free word order language of

[1] There is considerably variety in some aspects of the theory, a result of its constant evolving. An exposition, of the length given here, can at best present one instantiation of the theory. The one presented here was chosen because it is unencumbered by problem specific details making it easier to implement and present. For a more complete presentation of the theory see Haegeman (1991, 1994). Computational aspects are covered in Stabler (1992).

Australia). This concern with multiple languages has helped move Government-Binding (GB) away from the view of grammar as a set of rules toward the view that grammar is a set of interacting principles and parameters[2]. These principles determine how well formed an utterance is (*well-formedness*). This viewpoint has evolved from the realization that languages differ considerably at the surface level and would require very different sets of structural rules for their respective grammars. However, these disparate rule sets are generated from universally shared principles, and a few language dependent parameters. It is these universal principles that are the main concern of GB for they not only capture the attributes of all languages in an elegant and concise manner but provide greater explanatory depth than is possible with only a rule based, phrase structure grammar.

Theory and Principles

In the following, the core principles essential to GB based parsers are explained briefly along with some illustrations that show how the principles are used for parsing.

X-bar Theory: One of the principles of GB is an extremely simplified phrase structure component. This component, based on X-bar theory (Jackendoff 1972), posits that for all languages there is an underlying phrase structure template that can be used to adequately describe the language's structure. Most phrase structure grammars would require many hundreds, or even thousands, of language specific

2 This section will cover the main principles of GB but will not cover any of the parameters. The parameters are general in nature but are set for each specific language. Some examples of parameters are: the position of adjectives in relation to nouns (they come before the noun in English); whether subjects are optional or not in tensed phrases (optional in Spanish but required in French); and which phrase levels (noun phrase or sentence level) blocks certain principles such as bounding, which is discussed below.

rules for a generalized grammar. GB avoids this need by using two or three extremely general rules such as:

XP → Specifier XN Modifier.

XN → Preadjunct X Argument.

These rules are shown in template form in Figure 2.1.[3]

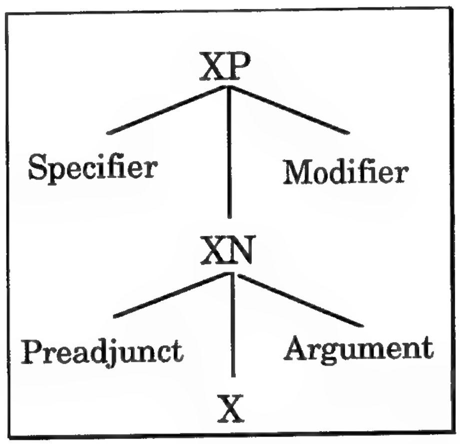

Figure 2.1. The X-bar Template

In this template, XP is any generalized phrasal expansion (ex., noun phrase, verb phrase), XN is an intermediary node and X is one of

- an "ordinary" lexical category (ex., noun, verb, or adjective)
- a complementizer (ex., that, which, who, what) or
- inflection (contains tense, agreement, modal verbs).

All the other slots in the template (i.e., specifier, preadjunct, modifier or argument) can contain zero or more phrasal or lexical items. For example: determiners and quantifiers are *specifiers* of noun phrases; adjective phrases and most adverbials

---

[3] The X-bar schema given is from Sells (1985). Jackendoff (1972) characterizes it as being binary (pg. 17) and three levels deep (pg. 53). The true form of X-bar is a matter of contention though most modern GB linguists prefer binary trees (Haegemann 1991). The choice of using Sells' model is mainly a circumstance of the development of our work. The implementaion of the parser, for the pilot study, was started before the restriction to binary branching was so widely accepted. We have incorporated that model (Sells') without modification.

are *preadjuncts*; noun phrases and prepositional phrases that are required by a verb are *arguments*; and additional prepositional phrases are *modifiers* of a verb.

Parsing within this framework involves instantiating the template for each phrase within a sentence. Example instantiations of the template for the sentence "A wise person eats salad" are given in Figure 2.2[4].

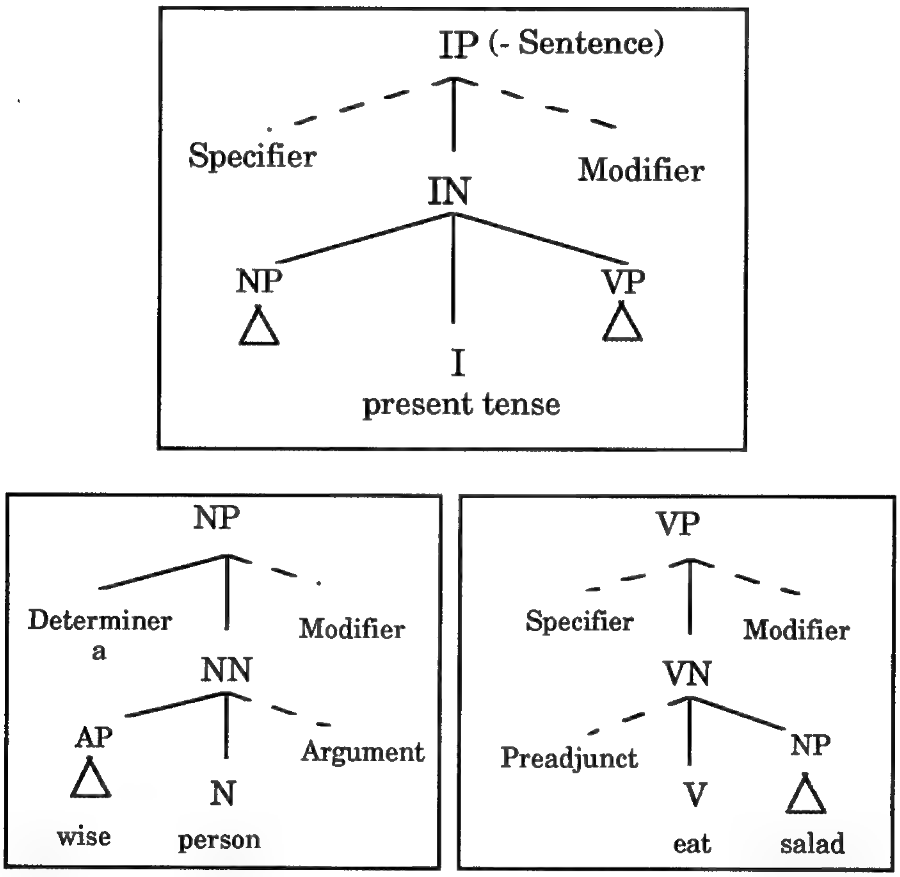

Figure 2.2. Sample Instantiations of the X-bar Template

The XP template is used to generate candidate parses, which are shown in the figure as *trees*. Most of the remaining principles can be used to limit the parser to

[4] The dashed lines in the diagram indicate empty slots in the template. The triangles represent internal structures of phrases (i.e., other instantiations of the template) which are left underspecified in the diagram as a matter of convenience.

accepting only grammatical strings. Since these principles are well-formedness rules for allowable structures, they can be used to reject the phrases that do not meet the standards of the principle.[5]

Following is a quick overview of the major principles.

Government Theory: Government is concerned with the relations between heads of phrases and their corresponding arguments. Government is defined in terms of a more primitive notion, *c-command* ('c' for constituent, i.e., a phrase). There are a number of ways to define c-command, each with its own nuances. The following is adapted from Sells (Sells 1985).

C-Command

α c-commands β if and only if every phrasal node dominating α dominates β.

Given this definition of c-command, government is simply a mutual c-command relationship between a head (i.e., an ordinary lexical category) and any other constituent. For instance, in Figure 2.2 the verb EAT governs the noun phrase SALAD because the verb phrasal node (VP) dominates (i.e., is above) both of them and a verb is always a head. In effect, government restricts itself to phrases, which are viewed as subsets of the parse tree.

Government is important because many of the other principles use government as a constraint for their own application. The determination of case, θ-roles, binding, and subcategorization must all be satisfied under the constraints of government.

[5] Remember that the X-bar template is a (very impoverished) phrase structure grammar. Therefore, a PBP need not rely on X-bar to generate its candidate trees. An (existing) phrase structure grammar could be used to generate the trees and the remaining principles could be used to filter out ill-formed candidates. This would still be correctly termed a PBP but one that would not share in X-bar's claim of syntactic coverage. The beauty of X-bar is that it will generate a candidate tree for virtually any sentence, phrase or fragment that it is given. It is extremely difficult to write specific PS-grammars that can claim this kind of coverage.

Move α: In the original formulation of transformational grammar (Chomsky 1957, 1965), which was the predecessor of Government-Binding theory, the idea of movement played a key role. Under that characterization, one sentence transformed into another by having a syntactic constituent "move". For instance, the sentence "Pat is a comic" could transform into "What is Pat?" by having the interrogative pronoun "what" replace the noun phrase "a comic" and move to the front of the sentence thus creating the question form of the second sentence. These transformations at the syntactic surface of a sentence are reflections of processes that are taking place in deep structure (the mental representation of the sentence, if you will). Movement was restricted by a multitude of constraints specifying what could and could not move. In the reformulation of transformational grammar to Government-Binding theory, movement became the principle *move* α. According to *move* α, "anything can move anywhere," the constraints on movement are simply the other principles and government.[6]

Bounding Theory: This is one of the principles that constrain movement. The constraints on movement, and related items, are couched in terms of bounding nodes that are parameterized for any given language. A bounding node is simply a node in the syntactic tree through which movement must take place. The bounding nodes for English are noun phrases (NP) and complementizer phrases (CP). Movement can cross more than one bounding node but must go through each one in turn. For example[7] in

[6] In the modern formulation, movement is viewed figuratively using *chains* (Correa 1988). Move-α, (along with X-bar's phrase structure characteristics) gives GB a mixed pedigree. While its phrase-structure, transformational roots still show, the modern formulation of GB is almost entirely principles and parameters based.

[7] Many of the examples of this section were either taken directly from or derived from Sells (1985).

the man who I think that you said that you had seen

WHO moves from the argument position of the verb SEE across two bounding nodes, marked by the THATs. If either of these nodes had been filled by a wh-phrase (ex., which, who, etc.) the movement would be blocked. This explains the contrast in the following pairs.[8]

(1) a. Pat said that this tool cuts wood.
    b. What did Pat say this tool cuts?

(2) a. Pat wondered who saw what.
    b. * What did pat wonder who saw?

In (1), WHAT moves through the bounding node marked by the THAT. In (2) the movement is blocked because the bounding node is filled with a WHO.

Binding Theory: This principle is concerned with relations of anaphors, pronouns, and names to possible antecedents. The antecedents are said to "bind" the anaphors and pronouns. It must be remembered that GB is a syntactic theory, and being such it is concerned with the syntactic phenomena of sentences *in isolation*. Therefore, pronouns that occur in a sentence need not be anaphors in that they may refer to some antecedent that occurs in a previous sentence. Extra-sentential anaphora is outside the scope of the theory. Again, the theory works under government or more specifically under a governing category. A governing category is the smallest NP or sentence that contains the element and a head that governs that element. According to binding theory:

---

[8] The '*' before the second example indicates ungrammaticality. Mixed or questionable sentences are marked by a '?'.

(Principle A) an anaphor is bound in its governing category;

(Principle B) a pronoun is free (i.e., unbound) in its governing category; and

(Principle C) a name (referential expression) is free.

This accounts for the following sentences.

(a) Mary recalled that John had painted himself.
*(b) Mary recalled that John had painted herself.
(c) Mary recalled that John had painted her.
?(d) Mary recalled that John had painted him.
?(e) He painted Max.

The governing category for each pronoun in (a)–(d) is the imbedded sentence beginning with JOHN. Sentence (a) is grammatical because the anaphoric pronoun, HIMSELF, is bound by JOHN in its governing category as it must be by principle A. Sentence (b) violates principle A and is ungrammatical. That is, MARY should be the antecedent of the anaphor HERSELF, but MARY is outside the governing category. This leaves the anaphor unbound. Sentence (c) fixes the problem by changing the anaphor HERSELF to the pronoun HER, thus principle A no longer applies and principle B is satisfied. Sentence (d) is ungrammatical under the reading that HIM refers to JOHN as it would violate principle B by having JOHN bind HIM. If HIM refers to someone other than JOHN then the sentence is grammatical as principle B is satisfied. A similar situation arises in sentence (e), this time with the name MAX. If MAX refers to HE then the sentence is ungrammatical as principle C has been violated, otherwise it is fine.

Case Theory: This principle stipulates that each overt noun must have one and only one grammatical case (e.g., nominative, accusative, ...). A further stipulation (a parameter for English) is that case must be assigned under adjacency. That is, only

nouns adjacent to their case assigner can receive case. This principle explains the contrast in the following sentences.

I like flowers very much
* I like very much flowers.

The first sentence is fine because FLOWERS receives accusative case from the verb. The second sentence is ungrammatical because FLOWERS cannot accept case from the verb due to the intervening phrase and it must receive case to be grammatical.

θ-Criterion: This principle refers to assignment of thematic roles for arguments of a verb and is covered in detail in the third section of this chapter.

The principles and theories outlined above, along with a few others, collectively form the grammar of all languages according to Government-Binding theory. Sentences are analyzed by first projecting a tentative parse structure using the X-bar template. Then the other principles are applied, under the constraints of government, to ensure that the final structure is valid. A structure that violates any of the principles is rejected as being ill-formed. Metaphorically, the principles can be viewed as a generator (X-bar) and a series of filters. See Figure 2.3.

Advantages of Government-Binding

There are a number of advantages to the Government-Binding approach. The first is conciseness and simplicity resulting from the sparsity of phrase structure rules. In contrast to the one template of GB, the Augmented Phrase Structure Grammar (Jensen 1986) of the PLNLP system contains hundreds of rules in the core grammar alone; additionally there are rules for parse ranking and parse fitting

(Fagan 1988). The creation of these rules not only requires a great deal of linguistic expertise but also generally requires some

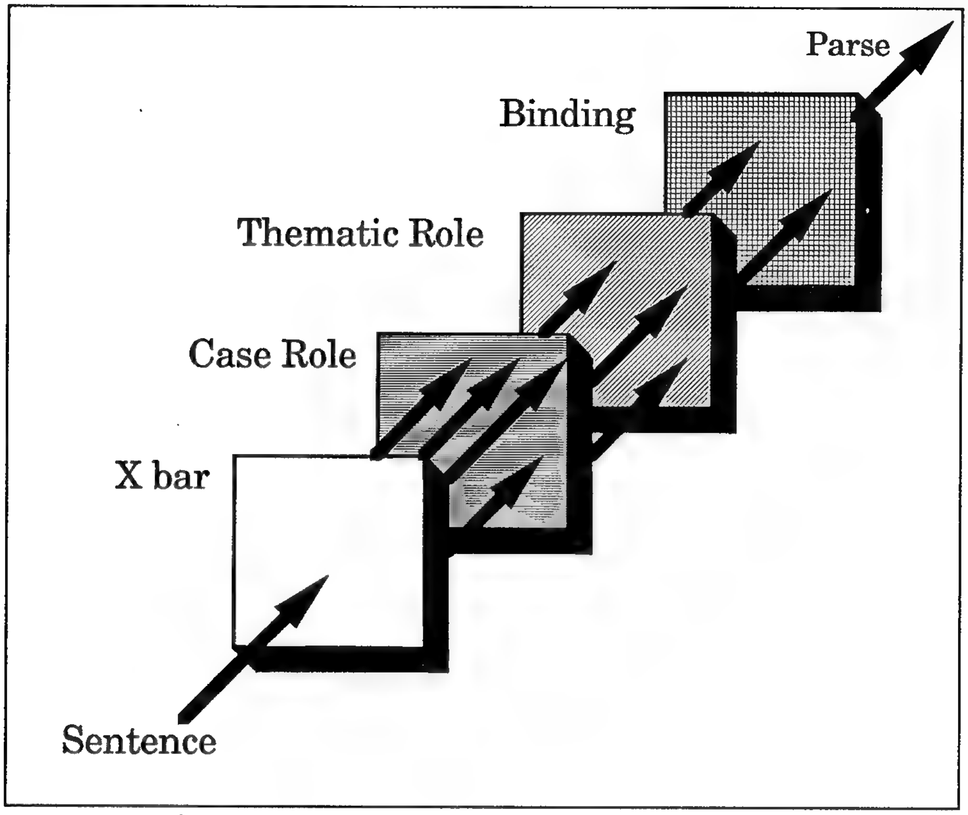

Figure 2.3. GB as Generator and Filters (adapted from Berwick and Fong 1990)

clever programming. Additionally, in rule based systems there is much redundancy between the rules and the lexicon. For instance, phrase structure rules alone are inadequate to determine correct attachment of arguments so the information must be represented in the lexicon as well as in the rules for phrase structure grammars. The following examples are demonstrative of lexical entry requirements for attachment.

put, Trans Verb, (*Agent*, NP), (*Theme*, NP), (*Locative*, PP)
eat, Trans Verb, (*Agent*, NP), (*Theme*, NP)

Since GB grammar is in essence a lexical grammar, this as well as the other information necessary to reconstruct a valid syntactic structure for the sentence would be in the lexicon. There would be no need to include separate rules in any other component (for instance, in the X-bar component). The lexicon for the current system will be covered in the next section of the chapter.

Another advantage of the GB approach over more general phrase structure approaches is in completeness of syntactic coverage. Most rule-based systems are developed incrementally against sample text so the most that is ever certain is that the rules cover the text already encountered. GB has been developed from general linguistic principles and not in response to any particular corpus of text. This comprehensive approach ensures complete coverage of grammatical text and allows graceful degradation for ungrammatical text. Instead of just failing for ungrammatical sentences, a GB system can indicate which module fails and thus return results that may be usable by a semantic or discourse component. By loosening the restrictions that the principles enforce, and incorporating a structured method of "relaxation rule," a GB parser can allow for ill-formed sentences and thus graceful degradation. For instance, the sentence given earlier, "I like very much flowers," violates case theory. Since LIKE cannot assign case yet FLOWERS needs to receive it, the assumption could be made that the two should be together. Thus a semantic interpretation could decide that the sentence was "I like flowers" with the intervening phrase either misplaced or of the wrong type.

A third advantage of GB is related to the second. GB is universal in nature and a correctly designed GB system is easily modifiable for other languages, including quite disparate languages (e.g., English and Japanese). To modify the system for a different language, the lexicon would have to be changed and a few language-specific parameters would have to be reset (cf., Dorr 1987). These parameters include whether case is assigned under adjacency, what parts of speech assign case, the location of modifiers in relation to the headword, and whether the language

requires pleonastic subjects (ex., **it** rained). In contrast to this, phrase structure grammars have to be painstakingly recreated for each language. The thousands of hours that were spent analyzing the first language would have to be respent in analyzing the surface structures of the second language and trying to cover these structures with a new grammar.

Another advantage of the principles and parameters approach is modularity and its impact on computational efficiency (Berwick and Weinberg 1984). The XP template is the main builder of candidate structures for the parser with most of the other principles acting to constrain the allowable types of structures. The order of application of the constraints (i.e., principles) is not a concern of theoretical importance but only of computational efficiency (Fong 1989, 1991). This means that each constraint can be built, debugged and maintained separately from the others, resulting in tremendous savings of effort. Furthermore, since any structure that fails any one of the constraints is ill-formed, it means that the candidate structures can be passed to the constraints in parallel, thereby improving efficiency. Efficiency is further increased, in comparison to other phrase structure grammars, by the size of the grammar. That is, increases in coverage are caused in GB by increasing only the lexicon, not by increasing the size of the grammar.[9] In most phrase structure grammars, either the number of phrase structure rules has to increase or the rules have to be re-written to give more syntactic coverage. This proliferation of rules is detrimental to efficiency.

To better understand the differences in the approaches of GB and general phrase structure grammars it may be easier to use an analogy. One could take two approaches to chemistry. The first would analyze all compounds by looking at the way the elements (carbon, hydrogen, etc.) combine to make a compound. That is, one could describe all compounds by knowing a few universal elements and the way

---

[9] Actually, the "size of the grammar" can be increased by decreasing the size of the parser, since X-bar overgenerates possible structures and the other principles filter them out.

in which they combine. The other approach would be to simply catalog all possible compounds. Then to describe a new compound one would compare it to something already in the catalog. The first approach (GB) results in an elegant and concise method that has broad explanatory capabilities. The second approach, while interesting to do, is extremely time consuming, explains relatively little and is never guaranteed to give satisfactory coverage.

Longman's Dictionary of Contemporary English

Longman's Dictionary of Contemporary English (LDOCE) provides the basis for the lexicon for the principle based parser being used for this research. LDOCE (1987) was designed for use by learners of English as a second language. It therefore demonstrates some differences from the ordinary English dictionary. A number of these differences are important to note.

The 55,000 or so words and phrases that are included in the dictionary were chosen for appropriateness as both core vocabulary and relevancy of current use. This choice of vocabulary results in the dictionary being a prime candidate for a basic NLP lexicon as there are fewer arcane or rare words than found in most dictionaries.

Additionally, the words are defined using a defining vocabulary of approximately 2000 basic words, thus the definitions are more easily understood. This limiting of the defining vocabulary, though, is a two-edged sword. On one side it cuts through the need to have a large vocabulary to understand the definitions. This would allow a computational system to "bootstrap" the definitions, that is, to work with the defining vocabulary and add other words as their definitions are processed. This approach would be very encouraging except that by limiting the vocabulary, many of the definitions require a more complicated syntax. This syntax

occurs especially in imbedded clauses that are in effect imbedded definitions. For instance, in the definition

computer an ELECTRONIC machine that can be supplied with a PROGRAM (= plan of operations) and can store and recall information, and perform various processes on it.

there is an imbedded definition of *program*.[10] This increase in syntactic complexity makes the processing of definitions more complicated than might be the case with other dictionaries.

Another feature of the dictionary is the inclusion of example sentences. There are over 75,000 example sentences, many culled from American and British newspapers. These sentences are designed to provide natural and typical examples of each word's usage. For a learner of the language, these example sentences provide an aid for learning the correct way to use a word. The sentences can also play this role for NLP systems by providing grammatically correct sentences on which to test the parser. If all the parameters are correctly set and each principle in the parser is working correctly, then the sentences should parse. Failure to do so would indicate a problem with the parser. This testing also provides the opportunity to discover the correct number of arguments for the verbs, and the nature of the semantic roles filled by the nouns. For instance, if the enhanced lexicon indicates that a word should have a θ-role of LOCATIVE and the example sentence does not contain a LOCATIVE then the θ-role would require examination.

The above information is in both the hard-copy and on-line versions of the dictionary. The on-line version, though, has information that is not present in the hard copy version. For instance, the verb *saddle* has an entry in the hard-copy version of LDOCE as:

---

[10] The words in capital letters in definitions in LDOCE indicate words that are not part of the defining vocabulary but are defined elsewhere in the dictionary.

saddle[2] /'sædl/ *v* [T (UP)] to put a saddle on (an animal): *He saddled (up) his horse and rode away.*

This entry provides the word, the pronunciation, the syntactic category, transitivity information (i.e., whether the verb requires an object), optional phrases (ex., UP), the definition and an example sentence. The on-line version has all of this information as well as some extra information. The subject field code lists this as EQ, an equestrian term. The box code lists a human subject (H) and an animate object (A). These *selectional restrictions* can be of great use to an NLP system by providing clues to the types of thematic roles a noun can support (for instance, inanimate objects cannot be AGENTs) and by limiting the semantic possibilities (to equestrian or figurative uses, for example). Additional information that is available for some entries includes box codes on country of origin, social register, level, period in which the word was used, language of origin, whether it is a new term, and if there are any cross references or illustrations.

Theory of Thematic Relations

The theory of thematic relations is based on work by Gruber (1965), Jackendoff (1972) and others (ex., Fillmore 1968, 1977, Chafe 1970, Anderson 1971). The theory is not completely unified as each version is slightly different from the others but they do share many common characteristics. The exposition given here is based on Cook (1989).

Thematic relations is a theory of semantics that views a sentence as a predicate with arguments. The verb is central (i.e., the predicate) to this approach, and the arguments have some role relative to the verb. These roles are known as thematic or theta (θ) roles. These roles can be viewed as a variant of Case Grammar (Cook

1989) and will be so viewed in this research[11]. These roles can be viewed as *who does what to whom.* An impetus for thematic relations was the recognition that two sentences with the same words but different structures can represent nearly the same meaning. For instance in the sentences

The goalie kicked the ball to the wing.
The ball was kicked to the wing by the goalie.

the subjects are different for each sentence but the meaning is essentially the same. The reason for this is that the roles for the arguments for *kick* in the sentences are identically filled, that is *goalie* is the AGENT, *ball* is the THEME, and *wing* is the RECIPIENT. That is to say, the semantic deep structures of the two sentences are identical even though the syntax is quite different.

Also of concern to the proponents of this theory is the universality of the roles. That is, the roles used should be representative of all languages. This concern, in conjunction with an inclination toward elegance, has limited the number of roles to generally less than a dozen. Cook uses just five: AGENT, THEME[12], EXPERIENCER, BENEFACTIVE and LOCATIVE. The roles that a verb takes as arguments are important, but Cook realizes that whether a verb is a *state, process* or *action* verb is just as important. This classification of verbs along two scales allows him to develop a matrix of verb types. For instance, *give* is an example of an ACTION-BENEFACTIVE

---

[11] In fact, though, the term Case Grammar will not be used again in this research even in places where it would seem more natural (e.g., in discussions of Cook's Case Grammar Matrix Model). The reason for this is not based on concerns of theoretical purity or parochialism, rather the term Case will be restricted to syntactic case (ex., accusative, dative) as is the custom in Government-Binding literature.

[12] Cook actually uses OBJECT but as was mentioned earlier *object* in GB is syntactic not semantic. The term THEME will therefore be used throughout this section to avoid confusion. This choice does not appear to be in opposition of Cook's given his own definition of OBJECT (pg. 191). "Object is the neutral underlying theme of the state, process or action described by the verb." Emphasis is mine.

verb. That is, someone (the AGENT) *gives* something (the THEME) to someone (the BENEFACTOR). The matrix model is given in Figure 2.4. The roles are presented using the first letter of the role (ex., A = AGENT, E = EXPERIENCER, ...), and Ts is a stative THEME. The boxes in the matrix each contain two frames and example verbs. For instance the verb *elect* has a frame of A,T,T as in "*They* (AGENT) *elected Jean* (THEME) *president* (THEME)." *Kill* has a frame of A,T. There are no verbs in English that have the frame T,B.

Cook also gives five rules to use to develop the frames (Cook 1989: 193).

(1) Each frame consists of a verb and one, two or three roles.
(2) The THEME role is obligatory to every frame.
(3) The secondary roles EXPERIENCER, BENEFACTIVE and LOCATIVE are mutually exclusive.
(4) No role except THEME occurs more than once in a frame.
(5) Roles are listed left-to-right by subject choice hierarchy.

There are some interesting points to consider here. Both rules 1 and 2 ensure that the deep structure for each sentence has at least one argument (a THEME). That means for sentences like, "It is raining." there is one role, the THEME. Since the *it* in this sentence has no real role to fulfill, the role is deleted at surface structure. Deletion of roles is common, as are optional roles.

It should also be noted that Cook does not use an INSTRUMENT role. He feels that all occurrences of INSTRUMENT are optional and are introduced by the prepositions in the sentence not by the verb.

Verb Types	Basic	Experiential	Benefactive	Locative
1. State	Ts be tall Ts, Ts be + N	E, Ts like Ts, E be boring	B, Ts have Ts, B belong to	Ts, L be in L, Ts contain
2. Process	T die T, T become	E, T enjoy T, E amuse	B, T acquire T, B ...	T, L move, iv L, T leak
3. Action	A, T kill A, T, T elect	A, E, T say A, T, E amuse (agt)	A, B, T give A, T, B blame	A, T, L put A, L, T fill

Figure 2.4 Case Grammar Matrix Model
(Adapted from Cook 1989)

The secondary roles EXPERIENCER, BENEFACTIVE and LOCATIVE are mutually exclusive (Rule 3) because they establish different semantic domains. For example, EXPERIENCER verbs deal only with cognitive processes (ex., *fear, hear, love, understand*) and therefore have no need to require locations as arguments. Likewise, there are no verbs with duplicate roles other than THEME as no verbs require two AGENTs, under Cook's analysis, or require an argument to be in two places at the same time (i.e., two LOCATIVEs).

With this fairly simple semantic scheme it is possible to create frames for all the verbs of a language. Moreover, Cook states that his roles are both necessary and sufficient (Cook 1989:192). This claim stems from both his analysis of all the other

major thematic role systems and by manually testing his system on actual text[13]. He also compares his system of roles with other semantic systems including the hierarchy used by Peter Roget. One nice result that thematic relations give is that when analyzed, verbs with the same frame cluster into semantically related groups and semantically related verbs have similar frames.

Government-Binding theory does not stipulate the number or type of relations that are allowable. GB does, however, provide a mechanism for ensuring that the relations are rigorously assigned based both on government and the subcategorization requirements of the verb. As an example, consider the phrase *in the car* in the following:

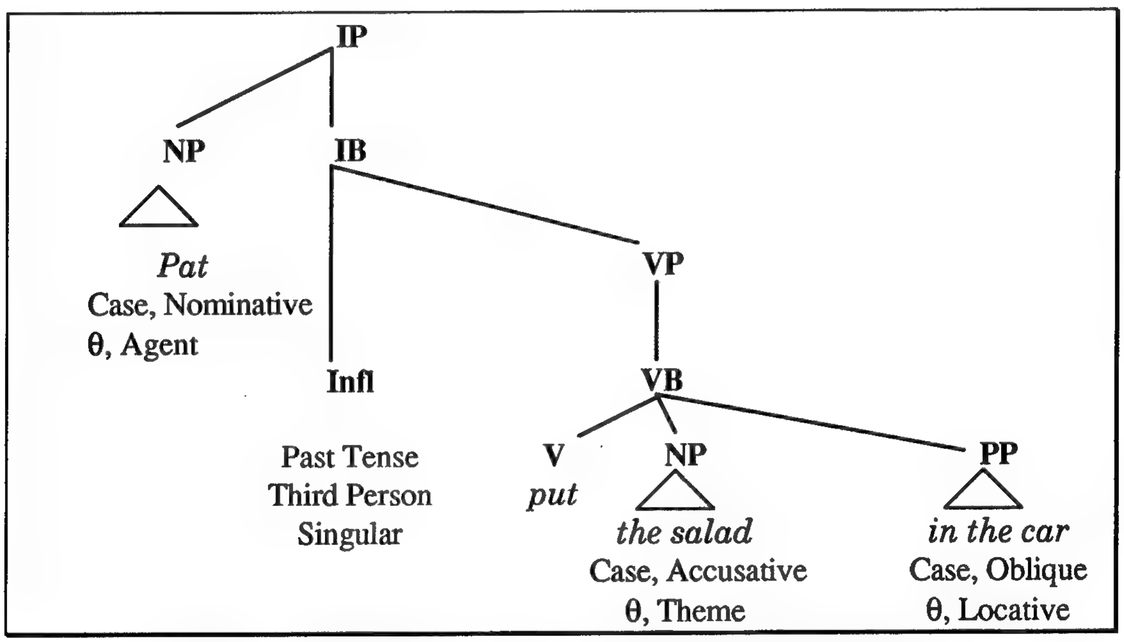

Figure 2.5. Parse of *Pat put the salad in the car*

Pat put the salad *in the car*.
Pat ate the salad *in the car*.

[13] This claim of necessity and sufficiency remains somewhat suspect. Sowa (1984) used just one basic relation (LINK) to develop a much richer set of relations than Cook uses and Zubizaretta (1987) and others in the GB literature question the need to label the relations at all.

In the first sentence (Figure 2.5) the verb *put* requires (i.e., subcategorizes for)[14] a direct object and a locative prepositional phrase so the phrase *in the car* has to be attached to the verb as the LOCATIVE of the object. The AGENT role of *put* would be assigned to the specifier of the INFL Phrase (the subject of the sentence), and the THEME and LOCATIVE roles to the siblings of the verb. All of this occurs under government. Since *put* subcategorizes for a locative prepositional phrase it must assign a thematic role of LOCATIVE to that phrase.

In the second sentence (Figure 2.6), *eat* does not require a locative and thus the prepositional phrase has to be attached to the sentence as the locative of the subject. The AGENT of *eat* would be assigned to the subject of the sentence, the THEME to the sibling of the verb, and the prepositional phrase would be attached to the verb phrase and not the verb itself. In this instance the verb does not subcategorize for a locative and the prepositional phrase receives no thematic role. This is a major difference between case theory and θ-theory. In case theory, every overt noun must receive case. While in θ-theory, only arguments need receive a thematic role.

With this form of lexical semantics presenting a tight, feasible alternative for NLP, one might assume that it could provide the semantic component for an information retrieval system. Unfortunately, it seems inadequate for that purpose (Lu 1990). The problem is one of richness.

---

[14] A good deal of overlap exists between subcategorization and the θ-criterion. Subcategorization is concerned with the syntactic categories that a verb requires while the θ-criterion is concerned with their semantic nature. How much of one is derivable from the other and how much redundancy is required in the lexicon is a matter of contemporary research.

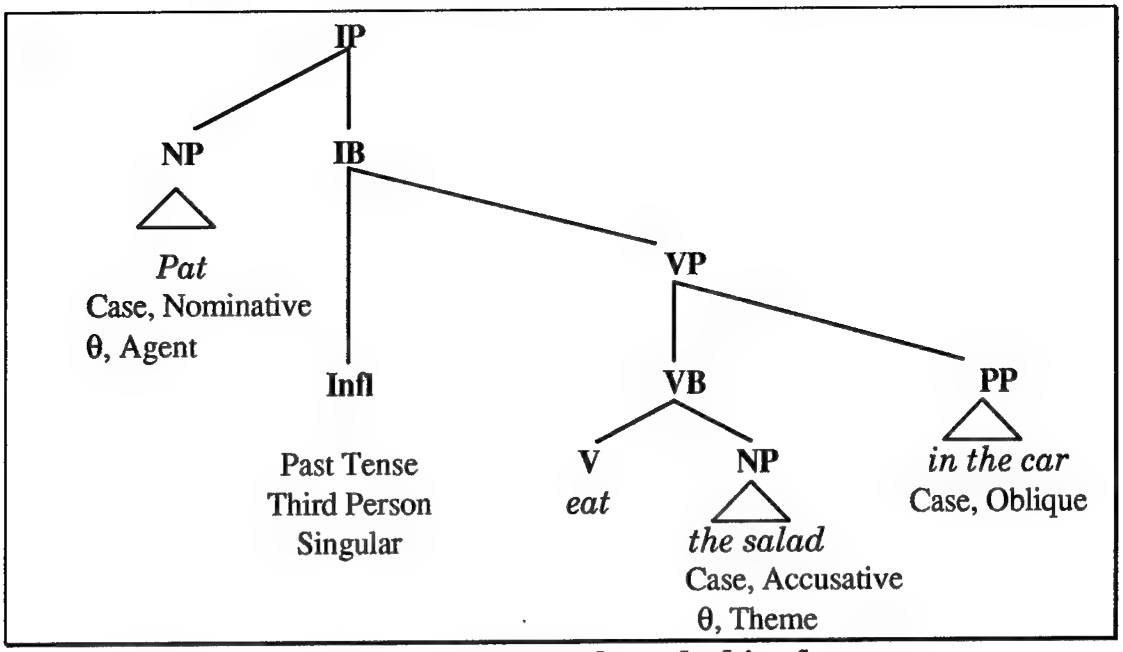

Figure 2.6. Parse of *Pat ate the salad in the car*

Not only do natural languages allow one to use different structures to make the same statement, they also allow the use of different vocabulary (i.e., synonyms). Thematic relations retain the verb as an essential element in the analysis, making the equation of two identical ideas difficult. For instance, the sentence, "Jean imbibed a few margaritas" might be analyzed as

imbibe(Jean/AGENT, margarita/THEME).

Thus searches using the keyword *imbibe* would find the statement whereas searches using the keyword *drink* would not. This obviously has limitations for information retrieval. The traditional IR method of handling this situation is to use a thesaurus (a list of synonyms) and search on synonyms. As was seen earlier, words are synonymous only on one or two senses. Therefore to correctly substitute synonyms in the search it is necessary to identify the sense of the word being used and use only those synonyms with the same sense. This is a difficult task to do automatically. What is needed for IR and similar purposes, then, is a richer

semantics than θ-roles can provide. One of the goals of this research is to show an approach to providing that richer semantics by projecting θ-roles into the conceptual hierarchy in *Roget's International Thesaurus*.

Roget's International Thesaurus

Roget's International Thesaurus (Roget 1977) is a rich, culturally validated source of information. Roget's original intent for the thesaurus was to provide a "grouping of words according to ideas" and not to produce a list of synonyms. This intent carries into current editions. Roget's fourth edition partitions the world of ideas into eight classes: *abstract relations, space, physics, matter, sensation, intellect, volition* and *affections* (Figure 2.7).

Each class is further divided into anywhere from 3 to 10 subclasses. For instance, class five, *sensation*, is divided into 6 subclasses: *sensation in general* and the five senses (*touch, taste, smell, sight, hearing*). Each subclass is divided into from 0 to 14 headings (ex., *touch* has zero, *hearing* has five). Each heading is divided directly into categories; the traditional entry point into the thesaurus. Each category is a basic semantic concept such as *odor, fragrance* or *stench*. Categories are divided into paragraphs that are grouped mainly by part of speech. Paragraphs are divided into semi-colon groups that contain the individual words. There are 1042 categories and approximately 200,000 words in the fourth edition.

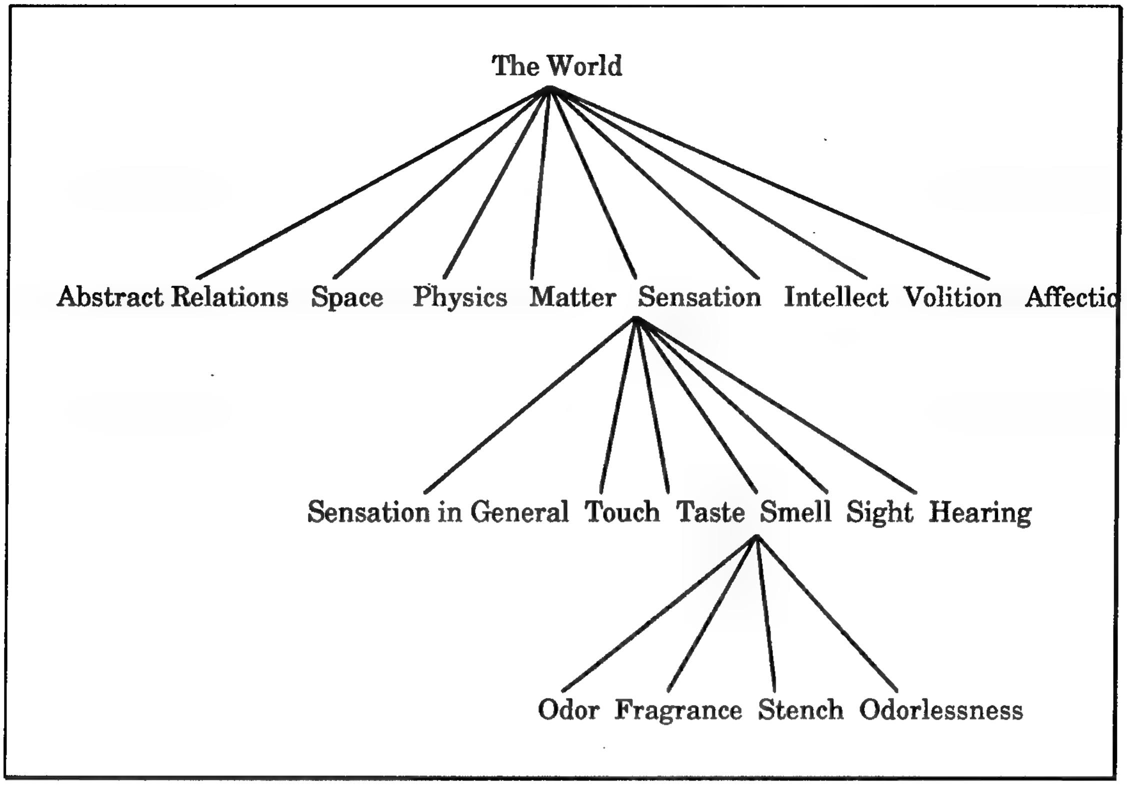

Figure 2.7. The conceptual hierarchy in Roget's

One on-line version of Roget's (Sedelow 1986) is a mathematical model of the index with some added enhancements. Unlike the hardcopy edition the on-line version is not arranged in hierarchical order. This limits its capabilities for some uses, such as browsing, as related words are not juxtaposed. The mathematical model makes it more useful in other ways though, as it has been optimized for computational processing by converting much of the structural information to ordinal values.

The size of Roget's makes it attractive as an NLP lexical source but even more attractive than mere size is the partitioning of the semantic space. By partitioning the world into classes, subclasses, headings and categories Roget has created a useful model of the cognitive world. Whether that model is "correct" or cognitively

valid is not a concern here.[15] It is assumed that the model is as correct as any other. Our world is constantly changing as are our viewpoints of the world. While no static model can hope to be valid for more than a few points in time, Roget's model has been used by English speaking peoples for 140 years and therefore the model may have become part of our linguistic culture. If that is so then Roget's model may be more valid than many other semantic models. The point is not really a concern in this research as the emphasis here is not on cognitive validity but rather it is on usefulness for computational processing, and this will be determined empirically.

As was shown earlier in the section on thematic relations it is possible, perhaps automatically, to generate θ-role frames for verbs. There is only a total of 24 possible θ-role frames in Cook's Matrix Model[16] which means that all the verbs of English must map to those 24 frames. This does not mean that there are only 24 semantic primitives in the language. Instead, each possible θ-role frame contains a number of clusters of related word senses. For instance, *kill, murder* and *stab* all have the same frame, as do *kiss, touch* and *hug*. While all six words are not closely related semantically there are two clusters of related words, which can be manually identified. These types of clusters usually inhabit only one or two categories in Roget's. It is this mapping of the words of LDOCE to the semantic space in Roget's that promises to be useful for NLP in general and information retrieval in particular.

---

[15] There is some evidence that our cognitive representation is in fact hierarchical (Miller 1991) though there are probably separate hierarchies for nouns and verbs.

[16] Actually Cook uses 33 frames. The extra ones are variations produced by an essential time role and by lexicalized roles (i.e., words that incorporate the role into their meaning and therefore the role is not present at the surface).

Summary

This chapter provided background on the two lexical resources being used. *Longman's Dictionary of Contemporary English* was chosen for this research for a number of reasons not the least of which is the richness of lexical information available in the on-line version. Some of the information that is required by the system, and described in chapter four, is not present in other MRDs. For instance, selectional restrictions are explicitly available as box codes in LDOCE but are unavailable in *Webster's Seventh New Collegiate Dictionary*. Other reasons for choosing LDOCE are less theoretical but nonetheless important. These include availability, researcher familiarity and the availability of support from other researchers in the NLP community.

Roget's International Thesaurus was chosen for very similar reasons. Other computational "thesauri" generally lack a conceptual hierarchy. They are, in effect, collections of synonyms and antonyms. The conceptual hierarchy makes Roget's fairly unique in that respect. Additionally, Roget's is much larger than LDOCE so that most of the words found in LDOCE are also found in Roget's.

This chapter also gave a summary of the two theories being used, Government-Binding theory (the theoretical foundation for principle based parsing) and the theory of thematic relations (the foundation for θ-roles and the starting point for the semantic component). The reasons for their selection have already been thoroughly discussed.

Chapter Three will continue with a discussion of some research that has used these components.

Chapter 3
Review of the Relevant Literature

The ancient masters were subtle, mysterious, profound, perceptive. The depth of their knowledge is unfathomable.

Lao Tzu, *Tao Te Ching*

This research brings together three main components, principle based parsing, machine readable lexical resources, and the theory of thematic relations, into an integrated system to improve natural language processing (NLP) capabilities for information retrieval (IR). Restricting the coverage of relevant literature to only those systems of immediate impact on this research would still result in scores of systems. Any coverage of that many systems would be superficial at best. Instead, this chapter will cover a few of the most apropos systems in some detail.

The first part will provide some background by discussing two contemporary systems that use NLP for information retrieval. Fagan's system relies heavily on syntactic analysis to select indexing terms to be used in the retrieval of whole documents. This use of syntax is common to the syntactic based approach (ex., Dillon and Gray 1983, DeFude 1986, Smeaton 1986, 1989, Fagan 1987, Salton, Buckley and Smith 1990) and has provided incremental improvements to information retrieval. The other use of syntax, which matches the syntactic parse tree of the query against that of part of the document (ex., Metzler and Haas 1989, 1990), has shown only mixed results. The use of semantics in conjunction with syntax to store and retrieve pertinent information in answer to queries, though, has been shown to be useful (ex., Berrut 1990, Jacobs and Rau 1990). This second system, the SCISOR system of Jacobs and Rau, demonstrates the power of semantics when applied to the retrieval of information in limited domains. It also supplies the impetus for the current research regarding the development of domain independent semantic processing and therefore will be covered in detail below.

In an integrated system, domain independent semantics requires domain independent syntax. As was shown in chapter two, Government Binding theory, the theoretical basis of principle based parsing, provides that domain independent syntactic coverage. The implementations of principle based parsers, however, have been limited to mainly theoretical investigations. This chapter will briefly discuss a few of these.

The work on thematic relations has been similar to that of GB in that most of the work has been theoretical in nature with few of the implementations being very germane to the present research. These works will also be briefly discussed except for the work by Chaffin and Herrmann and that of Sabah and Vilnat. Both pairs of researchers argue for a hierarchical representation for relations, and the works by them covered in this chapter were germinal to the present research.

The chapter will conclude with a brief discussion of two works dealing with machine readable resources. The work of Ahlswede dealing with the extraction of syntactic and semantic information from a machine readable dictionary is followed by a discussion of the research by Morris and Hirst that uses *Roget's International Thesaurus* to indicate the structure of text.

Natural Language Processing and Information Retrieval

NLP and IR have had a long relationship. This relationship has not been based on the brilliant successes of NLP for IR (because there have not been many) but rather on the continued belief of researchers that once research in NLP reaches an appropriate level of competence then it will greatly benefit IR.[1]

Syntactic Based Retrieval

Fagan (1987) compared the effectiveness of syntactic and non-syntactic methods for automatic phrase indexing for document retrieval. To do this, retrieval tests were run on five collections of abstracts. Some of these test collections contain both abstracts and queries, plus a list for each query of those abstracts in the collection that have been judged to be relevant. The procedures used were state of the art for

[1] See Mc Hale (1991a) for a review of the literature.

both the syntactic and non-syntactic methods. It must be remembered though that the intent of the study was to make the comparison using standard phrase descriptors (mostly noun phrases) and therefore the syntactic approach was limited in scope to those descriptors.

The non-syntactic approach used the proximity of co-occurring words as the basis of term selection. Rather than hard coding the degree of proximity needed, seven parameters were introduced: the **domain** (i.e., document, paragraph or sentence); the **proximity** (the distance between words within the domain); **df-phrase** (assures that the document frequency of the term is neither exceedingly high nor exceedingly low); **df-head** (identifies heads of phrases based on document frequency); **df-comp** (identifies phrase components based on document frequency); **df-st** (document frequency threshold for single term descriptors); and **length** (the length of the descriptor phrase). The settings for these parameters were optimized through extensive empirical testing for each document set. Additionally, each term was weighted and normalized to fit into the vector space model of the SMART system (Buckley 1985). A comparison was then made between the retrieval effectiveness of single-term descriptors (length = 1) and phrasal descriptors (length = 2). In those cases that the phrasal descriptors were less effective, a failure analysis was performed. The results indicated that there was a statistically significant improvement with the use of phrasal descriptors over single term descriptors and in those cases where single term descriptors did better it could be explained by the fact that the phrasal descriptors were selected without regard to word order or syntax, or because of over normalization of the phrases through stemming and ignoring stopwords. Most of these shortcomings appear to be solvable with syntactic methods of phrasal extraction.

The syntactic system used was PLNLP (Heidorn 1972). PLNLP (Programming Language for Natural Language Processing) has a broad-coverage grammar of English (Jensen 1986) based on an Augmented Phrase Structure Grammar

containing some hundreds of rules. It is capable of handling a wide variety of texts because it has a broad coverage of English syntax, it uses no semantic information to do its processing, and it has a base vocabulary of 130,000 words. Besides this grammar for well formed text, PLNLP has ranking rules for sentences that are syntactically ambiguous and thus produce more than one parse. If the sentence, or fragment, is not well formed then some of the grammar rules can be relaxed and a parse "fitted" to the sentence, thus producing an output that is hopefully useful.

This system was used to select phrase descriptors by parsing the noun phrases in the texts and using only nouns and their syntactic modifiers as descriptors. Much effort was spent ensuring that most of the allowable syntactic variations (i.e., adjectives, conjoined adjectives or noun phrases, adverbials, prepositions and conjoined nouns within prepositional phrases) were covered. Syntactic phrasal descriptors were optimized against the texts for this part using four parameters: **parse threshold** (determines the number of parses from which descriptors are drawn); **query parsing mode** (whether query is a sentence or noun phrase); **phrase subvector weight** and **df-phrase** (assures that the document frequency of the term is neither exceedingly high nor exceedingly low).

Again the phrase descriptors were tested against single-term descriptors but in this case for only two of the test sets. These were the sets on which the non-syntactic approach did the best and the worst. The syntactic approach has comparable results doing well on one set (statistically significant) and not showing much improvement on the other.

When comparing the two approaches, **the optimized non-syntactic approach seems to do better overall than the syntactic approach**. There are a number of reasons for this and Fagan goes into some detail studying some of them. Three of the reasons mentioned were: (1) the "queries" were for a large part already largely made up of key words and noun phrases and therefore were already syntactically processed by the users. (2) Even though PLNLP is capable of producing parses for

the full text, this study was limited to only the noun phrases and nothing else. This limited some obvious advantages that might be realized from a pure syntactic approach. (3) The selection of descriptors was based exclusively on syntax and did not include other areas of NLP (ex., morphology, semantics, pragmatics) that would probably fill some of the shortcomings of syntax alone. One point that Fagan does not make is that the "non-syntactic" approach identified sentences, heads and components of phrases by using statistical means. While this does not constitute syntactic parsing it may be inappropriate to label it as "non-syntactic," perhaps non-linguistic might be a more apt descriptor.

In related work, Salton et al. (Salton, Buckley and Smith 1990) compared syntactic and non-syntactic retrieval methods applied to two chapters of one of Salton's books. The syntactic method again used in this comparison was restricted to just noun phrases. The results showed comparable precision for phrases and "that the recall output is somewhat higher for the syntactic output than the statistical." Their conclusion, based on the amount of processing required, was that "the syntactic methodology is not recommended at the present time for automatic term phrase generation systems."

Fagan's system restricts its use of NLP techniques to syntax. This is done because full semantic analysis of unrestricted text is beyond current capabilities. By restricting the domain, though, semantics becomes a more feasible tool. The following system uses semantics, as well as morphology and syntax, in restricted domains.

Semantic Based Retrieval

SCISOR (System for Conceptual Information Summarization, Organization and Retrieval) is a prototype retrieval system that monitors news wire sources for

articles about corporate mergers and acquisitions (Jacobs and Rau 1990). As it identifies the pertinent articles, it also identifies the *target*, the *suitor* and the *price* tendered. Once this information has been acquired the system stores it in its knowledge base and can then answer questions pertaining to the articles.

To identify those articles that are likely to be pertinent SCISOR uses a variety of approaches that range from doing keyword searches on headlines to semantic analysis of text. This range of approaches acts as a set of increasingly sophisticated filters that select candidates for deeper analysis. One of the key features of these filters is the ability to identify those articles in which the system is not interested. This is an important feature as approximately 70% of the articles can be discarded through examining the headlines alone. Those articles that a filter identifies as pertinent receive deeper analysis while those articles that a filter is unable to classify are sent to the next filter. The first few filters are keyword-based but the system also uses two NLP parsers for more in-depth analysis. The first parser, a top-down skimmer relies on structural expectancies to help tag the key players. For instance, if the article is about two companies and one is the suitor then the other is probably (i.e., expected to be) the target. The other parser works bottom-up from the text and is used when more detailed processing is necessary, though the results may be only a partial parse. All the filters work together to supply the semantic component with the conceptual details obtained by the system which are stored in a semantic net. Using this representation SCISOR is able to directly answer user queries with the information that it has processed and acquired.

Much of the capability of the system is possible because it is operating in a very restrictive domain. The size of the domain helps in a number of ways. The size limits the parts of speech that a word can have, as previously mentioned. For instance in the SCISOR domain the word *tender* is a verb and not an adjective or noun. It also limits the semantic senses of a word. In an unrestricted domain the word *target* can have a number of meanings (ex., goal, limited group, etc.) but in

this domain it refers to the company that is being acquired. Also to be considered is the limited structure of the articles. As mentioned above this structure can be exploited by using an expectation based parser. The overall result is an extremely viable NLP system for the domain.

One problem that Jacobs and Rau recognize is in the evaluation of system performance (Jacobs and Rau 1990:96). Their system is so different from the standard IR system that meaningful comparisons are difficult. In one test they ran on a full day's traffic (729 articles) they were able to correctly identify slightly better than 90% of the stories on mergers and acquisitions. For these articles SCISOR correctly identified the target and suitor 90% of the time, the price-per-share 79% of the time and the total value of the offer 82% of the time.

It should be noted just how different SCISOR is from a traditional IR (i.e., document retrieval) system. The type of query one could imagine using with SCISOR might be, "What high-tech companies were subject to hostile takeovers last month?" Trying to phrase this query in such a way that a traditional IR system could find an answer without returning thousands of the 20,000 candidate documents would be difficult, but it is precisely this kind of conceptual retrieval that is the aim of semantically based IR systems.

SCISOR is a prototype information acquisition and retrieval system that uses a variety of methods including syntactic and domain dependent semantic processing. This reliance on domain characteristics does not prevent the system from being ported to other domains. SCISOR, in fact, has been ported to a number of other domains and has demonstrated its usefulness in these domains. SCISOR's power lies in its domain dependent approach and the modular way it is written which allows portability. Its domain dependence is also its main limitation as the system is portable but not generalizable. Until the capability to handle domain independent semantics is obtained, the information retrieval applications that use NLP techniques will remain severely limited in size.

Principle-Based Parsing

The work on principle based parsing (PBP) has been mainly theoretical in nature. These systems have tested the computability of the underlying Government-Binding (GB) theory (ex., Abney and Cole 1985, Berwick and Fong 1990, Correa 1988) or demonstrated the benefits of the GB approach to translation (ex., Sharp 1985, Dorr 1987, Wehrli 1990). Most of the PBP implementations have restricted their lexicons to cover small domains so there has been little serious attempt to demonstrate domain independence. Lin (1994) is an exception, deriving the lexicon for the system from two large on-line dictionaries. The system only includes syntactic information from the dictionaries (i.e., subcategorization features) and thus is of limited value to the current work. The works of Dorr (1987) and Fong (1994) are especially useful in constructing a GB parser but their systems were intentionally limited to small subsets of English, Spanish and other languages. The work of St. Dizier is a good example of an investigation of the computability of GB theory. St. Dizier (1989) uses Constraint Logic Programming (CLP) for principle based parsing to control the principle interactions. This type of control becomes especially relevant when PBPs are used with large lexicons. St. Dizier's interests lie in CLP and in particular in investigating the exploitation of CLP for natural language processing. His 1989 paper deals with controlling movement in GB. As was shown in chapter two, movement in GB is an invocation of the principle *Move-α*, which allows any constituent to move anywhere in a sentence as long as the other principles are not violated. If *Move-α* were implemented directly the resulting unconstrained movement would allow a plethora of possible parse trees, most of them invalid, which would then have to be filtered out by the other principles. By using CLP techniques St. Dizier is able to restrict the movement to valid trees.

The restriction of movement is somewhat metaphorically used here. No syntactic constituents in St. Dizier's system truly move. Instead, traces of the movement are placed in the surface level representation of the sentence and the principles work on the resulting *chains* (i.e., the traces and the constituents they represent). This approach was first championed in Correa (1988) and is the same approach that is being used in the present research.

Theory of Relations

The work on thematic relations has been even less concerned with computational implementations than that of general GB theory. While early work in Case Grammar did influence Conceptual Dependency theory (Schank 1976), which has been implemented (ex., Cullingford 1986), most of the work, in what can be strictly construed as thematic relations, has been theoretical in nature. Typical of this latter type of work is Volume 21 of the series *Syntax and Semantics* (Wilkins 1988). Of the thirteen papers in the volume most are either concerned with ramifications of the θ-criterion (i.e., the principle controlling θ-roles) as challenged by data from a particular language (ex., Choctaw, Russian, Warlpiri) or by difficulties arising from a given syntactic phenomenon (ex., reflexivization, passives). None of the papers are sufficiently relevant to the current research to demand an in-depth review.

During the same time that the θ-criterion was developing, a number of researchers were looking into the identification and use of different sets of relations for various purposes. Sowa (1984) provided a catalog of fundamental relations that he felt were useful in the analysis of text. The list was intended to be used, as needed, with his Conceptual Graph knowledge representation schema and was meant to be representative and not exhaustive. Dick (1991) combined some of Sowa's relations with those of Somers (1987) to provide a very domain-specific

knowledge representation for the retrieval of contract law cases. The system was not implemented but early results indicate that the knowledge representation does provide a reasonable approach to IR.

Relations as a Cognitive Hierarchy

Chaffin and Herrmann (1987) extracted 31 relations from the literature. These relations were then sorted by subjects and the results of the sorting were analyzed by cluster analysis. The result was a hierarchy consisting of five families of empirical taxonomy: *contrasts, similars, class inclusion, case relations* and *part-whole*. They were further able to show that the relations are decomposable into relation elements such as *dimensionality* and *intersection*. For instance *asymmetrical contraries* (ex., hot-cool, dry-moist) can be decomposed into *continuous bipolar dimensionality*. For the hot-cool pairing this can be understood by realizing that temperature is continuous (i.e., temperature can have any value above absolute zero) and wherever hot and cool are on the dimension of temperature they are on opposite sides of the midpoint (i.e., bipolar). This type of componential analysis has been used before for analyzing verbs (McCawley 1968), but it requires a rich lexicon. McCawley, for instance, analyzed *A killed B* as *A caused B to become not alive*. To arrive at this form using LDOCE one could first find the definition of *kill* which is *to cause death* and then look up the definition of *death* which is *the end of life*. From this *kill* could be defined as *to cause the end of life*, but to reach McCawley's form the inference that *the end of life* equals *not alive* would still have to be made. This is an inference that depends as much on world knowledge as it does on anything defined in the dictionary. To understand the amount of knowledge that is needed consider that *the end of X* implies *not in the state of X* only for certain types of *X* and only then in certain contexts. For instance, *the end of a pencil* does **not** imply *no longer a pencil*, and it seems most strange to

equate the sentence *"he was converted at the end of his life"* with *"he was converted when no longer alive."* This could be handled by having a good tense logic but it also indicates that a lexicon that is used for componential analysis would probably have to be mostly hand-coded and not derived from current machine readable dictionaries. Therefore componential analysis will not be used here.

Other Relational Hierarchies

Also using the hierarchical framework is the work of Sabah and Vilnat (1991). Their hierarchy is based on Case Grammar with CASE at the top and six cases (AGENT, ADDRESSEE, OBJECT, INSTRUMENTAL, SITUATIVE and DESCRIPTIVE) at the next level. The hierarchy has a total of four levels and has 30 cases at the bottom of the hierarchy. The relations were empirically derived by analyzing texts, natural language grammars and the literature on Case Grammars. These relations were then repeatedly grouped together to create the different levels of the hierarchy. This hierarchy was then used by a rule based parser (containing around 500 rules). The parser gives Sabah and Vilnat's system the capability of producing semantic representations of sentences with varying degrees of relational specificity. The result allows them to flexibly analyze the sentences in their domain. The difference between their work and the current research is two fold. First, transporting their system to a new domain would require some modification of the grammar, an artifact of a rule based grammar. The transportation may or may not be easily accomplished though they indicate that it would be relatively easy compared to other rule based systems. The current research is using a PBP which would require no changes to either the grammar or the parser. Transportation of a PBP requires resetting around a dozen parameters such as the ordering of the language (i.e., *subject, object, verb* vs. *subject, verb, object*). In both cases the lexicon must be replaced with one suitable for the new domain. Second, they hand built the

small hierarchy they use. The current research is using a hierarchy that is much more comprehensive and much more closely tied with the linguistic features of English. The initial implementation of Roget's thesaurus took over three years to produce and the thesaurus and its hierarchy has been constantly updated and revised over the last 140 years. Thus Roget's represents a sizable, lexical resource that few NLP practitioners could attempt to duplicate. While the resources being used in this research differ from the above researchers', the influence of Chaffin & Herrmann and Sabah & Vilnat, on the present research, must be credited.

Machine Readable Resources

The chapter will conclude with a brief discussion of two works dealing with machine readable resources. The work of Ahlswede dealing with the extraction of syntactic and semantic information from a machine readable dictionary followed by a discussion of the research by Morris and Hirst that used *Roget's International Thesaurus*.

The selection of papers for this section was difficult given the increased interest in computational lexicography for NLP over the last few years.

Guo (1989) worked on making an earlier version of LDOCE usable as a lexicon. The lexicon was developed by completely hand-coding 1200 KDV (key defining vocabulary) entries and then bootstrapping the rest of the defining vocabulary. That approach is excessive to the needs of the present research.

Carroll and Grover (1989) also discussed the extraction of a lexicon from LDOCE. Their approach is quite different from Guo's in that they decided the extraction cannot be made completely automatic. The bulk of their report deals with the creation and functioning of the tools they developed to aid in the extraction. Their objections to complete automatic extraction were based on the need to produce a

finished NLP system. These objections, for instance a few inconsistencies in the grammar codes, are not a major concern of the present research.

Boguraev and Briscoe (1989) explains the extraction of the grammar codes from the earlier version of LDOCE. These codes represent the subcategorization features of the words. The later version of LDOCE has encoded them differently but Boguraev and Briscoe's approach is still very much apropos to the extraction. Their work, along with Akkerman (1989), provides a solid base for the understanding and extraction of the grammar codes.

The work of Byrd (Byrd et al. 1987) uses a large (100,000 word) lexicon derived from machine readable dictionaries. The lexicon was developed by integrating the information from a number of dictionaries and other lexical resources. As such, the scope of their work greatly exceeds that of the present research. While the present work will be using two lexical sources (LDOCE and Roget's), the emphasis is not on the integration of the knowledge in these sources but rather on mainly co-locating these resources.

Ahlswede's work was selected for more detailed coverage because the extraction methods he developed appear to be the most useful way to approach the extraction of θ-role frames from the dictionary. As will be seen below, he had a different goal for his work but the methods should be useful to other tasks as well.

Using On-Line Dictionaries

In his dissertation, Ahlswede (1988) investigated the use of two approaches in the analysis of dictionary definitions in *Webster's Seventh New Collegiate Dictionary* (W7). The first was an NLP approach that used Sager's Linguistic String Parser (Sager 1981) to completely parse the definitions. The second used the UNIX™ text processing utilities (ex., grep, sed, lex) to extract patterns from the definitions that were then interactively processed. Both approaches resulted in the

development of relational triples of the type *love* SYNONYM *like* or *car* HAS-PART *wheel*. The results were rather mixed.

The NLP component was labor intensive both from a human and machine viewpoint. The component took

> "... a man-year or so of development time for the definition grammar, during which considerable computer time was spent parsing batches of definitions and considerable human time spent poring over bad or failed parses, and rewriting the grammar." (Ahlswede 1988:151).

The considerable computer time turned out to be 180 hours of CPU time on a VAX 8300 to parse the 8,000 or so definitions. The average time per parse varied considerably depending on the syntactic category of the defined word. Adjective phrases took an average of 10.59 seconds per parse while transitive verbs took 48.33 seconds per parse. The parser had an overall success rate of around 70%. The reasons for the parser failing to parse a definition were extremely varied and Ahlswede felt that the minimal improvements expected from rewriting the grammar to improve the success rate would not be worth the amount of effort required.

In contrast to this, by using the text processing utilities approach he was able to generate 11,596 relational triples for the intransitive verb definitions in just three hours. The quality of the triples thus produced was comparable to those produced by parsing.

Viewed in this light the parsing seems to be wasted effort but Ahlswede does not think so. There seems to be two reasons for this. The first is that the parser that he used is quite slow by current standards. This is due in part to its being a rule based grammar that attempts a comprehensive coverage of English. There are numerous ways of speeding up the parser for this particular application. One of the more obvious ways would be to limit the grammar to only the part needed to cover the

linguistic variations found in the dictionary. This could be accomplished by selecting a random subset of definitions and parsing them, keeping track of which rules of the grammar were used. This approach, in essence, would be comparable to writing a grammar specifically for the dictionary but would be less work. Of course, speed in this case would not be optimal because of the considerable overhead associated with the sophisticated user interface of the parser. The second reason that Ahlswede thinks that parsing is worthwhile is of more theoretical importance. The parser was able to identify some relations that were not representable using the relation-triple grouping. For example, the definition of *dodecahedron* is *a solid having 12 plane faces*. The parser gives a syntactic representation for the complete definition but in triples the representation is limited to *dodecahedron IS-A solid*, or *dodecahedron HAS-ATTRIBUTE faces*. It cannot represent *dodecahedron HAS-ATTRIBUTE faces NUMBER 12*. Certainly *12* is not an attribute of *faces* but rather of *dodecahedron*, and then only when referring to the number of faces. This type of relation remains a problem for the simple relational-triple.

Ahlswede's dissertation presents many useful techniques and illuminates many interesting facets of machine-readable dictionaries that are of direct benefit to this research. In particular the analysis of W7, which lacks any explicit selectional restriction information, reinforced the selection of LDOCE for the current research (the information on selectional restrictions is useful for the θ-Criterion). The speed of parsing is only of minor concern to the present research and then only in terms of relative speed with different lexicons. The relational-triple representation is not being used for the present work.

MRDs have been exploited much more by NLP researchers than Roget's has, but still a choice had to be made, among the papers, for the purposes of this section. The work by Sedelow and Sedelow (and students) has been based on a mathematical model of the thesaurus, and the emphasis has been on exploiting that model using

mathematical methods. While the current research will involve mathematical algorithms for the correct placement of senses of words into the thesaural hierarchy, the intent of the research is much closer in spirit to the work of Morris and Hirst.

Using Roget's

The last work to be considered deals directly with *Roget's International Thesaurus*. Morris and Hirst (1991) developed a method of computing *lexical chains* of related words in text. A lexical chain is "a succession of a number of nearby related words spanning a topical unit of the text" (Morris and Hirst 1991:22). The existence and structure of these chains help to illuminate the cohesion and structure of the underlying text.

The chains were created by comparing the relatedness of the important words in the text to each other. The determination of relatedness is based on five rules.

1. The two words share a category in their index entries.
2. An entry in the first word's categories has a pointer to the other word.
3. One word is a label of the other word's index entry or is in a category of the other word.
4. The two words are in the same group.
5. The two words have categories in their index entries that both point to a common category.

If any of these rules applied, or if there is one transitive link between the words, then they are related. That is, if *a* is related to *b*, and *b* is related to *c*, then they consider *a* related to *c*. They also note "The chains were built by hand. Automation was not possible, for lack of a machine-readable copy of the thesaurus. Given a copy, implementation would clearly be straightforward." (op.cit., pg. 29). The use of transitive links and the assumption that implementation would be straightforward

leads to some problems that will be discussed later. First let us examine their results.

The article gives an in-depth example for one of the texts they used. The text was analyzed by hand and the resulting chains were compared with the structure produced using the theory outlined by Grosz and Sidner (1986). In particular they compared the chains with the *intentional structure*, that is, the topics that the writer of the article intended to discuss. The comparison clearly shows that the chains, which are computable, "provide a good clue for the determination of the intentional structure. In some cases, the chains correspond exactly to an intention." This is encouraging indeed for a number of reasons. First the computation of the structure of a text is important to the understanding of what a text is about. This type of process is necessary for robust NLP and should prove useful for IR. Second, while Grosz and Sidner do provide a method of discovering the structure of a text they give no indication how this method might be implemented. The correlation of chains to structure shows one way this might be done. Finally, the production of lexical chains in this manner demonstrates that domain independent processing of text is possible using large, machine-readable lexical resources, in particular, *Roget's International Thesaurus*. This last point is particularly encouraging in the present context.

Now that we have seen the highlights of Morris and Hirst's work, let us stamp their work with a caveat regarding the use of transitive links and the assumption of straightforward implementation given above. As was mentioned in chapter two, the on-line version of Roget's is really an enhanced copy of the index and not of the actual entries. What that means is that an entry in the on-line Roget's (ex., resonance) has the entry numbers that are associated with it (ex., 323.1, 454, 323.4) and not the actual words that are found at the entry (ex., oscillation, vibration, ...). This is an important distinction in that it is the words **at the entry number** (ex., periodicity is in the entry 323.1) that have pointers associated with them (ex.,

periodicity 137.2). The pointers are the way that is used to show which sense of a word is intended. These pointers are not in the on-line version but are required for the second rule of Morris and Hirst's determination of relations. Given this constraint, implementation would not be straightforward.

The question of transitive links is more fundamental than that of implementation. In the section discussing links Morris and Hirst give the chain {*cow, sheep, wool, scarf, boots, hat, snow*}. They say that obviously unlimited transitivity should not be allowed as *cow* is not intuitively related to *snow*, and in general two or more links causes the relationship to be non-intuitive. With this reasoning and because they needed one transitive link to compute their intuitively formed chains they argue that one link is sufficient. Even using their example, one link is one too many. How can one feel justified saying that *cow* is intuitively related to *wool*? Furthermore, what is not evident in the way the rules are presented is that the links must use the same sense of a word. For instance since *club* is related to *diamond* (playing cards) and *diamond* is related to *ruby* (gem stones) the inference that *club* is related to *ruby* could be made. This is obviously counter-intuitive. If the sense of *diamond* were the same then this type of example would be eliminated. An interesting question, but one that will not be addressed here, is given unconstrained linking, what is the minimum number of links required to "relate" any two random words. The chain of length seven above is probably representative.

The problems with the use of transitive links and the assumption of straightforward implementation should not detract to any great extent from Morris and Hirst's overall findings. The approach seems encouraging and most likely could be implemented with relatively minor changes. An actual implementation would require the polishing of the algorithm with extensive empirical testing. Such an effort should prove well worth while.

Summary

This chapter has covered some of the more interesting literature that is germane to the present research. In doing so it has attempted to weave a coherent, thematic web around the literature. It started with current NLP approaches to information retrieval. It was seen that syntactic analysis alone is not powerful enough for IR. The addition of semantics to NLP greatly increases its capabilities and indicates some of the benefits that can be obtained if the limitation of domain dependence can be overcome. The approach used in the present work to overcoming this limitation is built solidly on the cornerstones of principle based parsing and thematic relations. The literature on these two areas was covered briefly as the current research is not trying to further the areas in any theoretical sense, rather it is testing their practical limits. Richer relations than those usually used in thematic relations have been shown to improve both the precision and recall of IR systems (Wang et al. 1985). Sabah and Vilnat showed that a reasonable set of relations is derivable from thematic relations. The work of Chaffin and Herrmann demonstrated that relations are cognitively formed as a hierarchy; a hierarchy that in many ways is similar to Roget's. Obviously the best representations are useful for domain independent processing only when they contain enough general information to cross domains. That type of information is difficult to acquire in hand-coded lexicons, so the chapter continued with an examination of Ahlswede's work on extracting information from W7. Finally, the work of Morris and Hirst was shown to demonstrate some of the latent domain independent power of machine-readable lexical resources.

The next chapter will cover in detail the approach and methodology that the research is using to measure the impact of the MRD on the parser. The approach will logically follow from the background provided by this and the previous chapters.

Chapter 4
Impact of the Lexicon

As sheer casual reading-matter,
I still find the English dictionary
the most interesting book in our language.

A.J. Nock

Approach

As shown in chapter one, the extension of NLP to include domain independent semantics can prove useful to many fields both from within and outside Information Science. This research posits one way of providing that extension – a principle based parser (PBP) that uses a semantically enriched lexicon derived from machine-readable lexical sources. The overall impact of such an extension would be the production of a rudimentary NLP system with domain independent semantics.

The research question that is being considered is, *how much impact does a large, general lexicon have on a principle based parser?* Answering this question will be useful for judging whether using a PBP with a machine-readable lexicon is a profitable enough path to warrant further exploration in the attempt to reach the goal of domain independent semantics. This question involves the lexical extension of PBPs to domain independence.

The overall approach being taken is a systems approach. That is, system components were built to demonstrate that the approach works. This type of approach is well steeped in the traditions of both NLP and IR. For NLP the systems approach seems obvious. Natural language processing without a system seems almost an oxymoron, as *processing* presupposes a system of some kind. These systems are seen by NLP researchers as a necessary tool for the verification of their ideas. The same arguments hold for IR. Many "intuitive" ideas have been shown to be wishful thinking by the process of building a system and establishing the idea's utility (or lack thereof). A number of the systems produced for NLP and IR were covered in chapter three and will not be revisited here. The important points to be made here are that this approach is traditional from both the NLP and IR viewpoints, the approach provides a way to answer the research questions, and the questions would be difficult to answer using other methods.

Methodology

This question deals with the impact of the lexicon on a PBP. The question was tested by measuring the performance of a partial PBP on a single corpus while varying the lexicon. First, a small, hand-coded lexicon was used with the PBP and

Task	Test	Output	Reason for Test
Develop hand-coded lexicon and X-bar parser with subcategorization	Run against test corpus	Number of structures	Create baseline
Replace hand-coded lexicon with LDOCE derived lexicon	Re-run parser against corpus	Number of structures	Assess impact of larger lexicon

Figure 4.1 Methodology

tested. This produced a baseline for results. Then, the hand-coded lexicon was replaced with one derived from LDOCE. The testing of this lexicon produced the results needed to show the impact of the larger lexicon on the PBP. This chapter goes into detail on both the methodology and results.

Partial PBP

The quest for domain independent NLP processing starts with the syntactic parser. For this component a PBP, based on Chomsky's Government-Binding theory, (Chomsky 1981, 1982, 1986) was used. The choice of using a PBP for this research is motivated by a PBP's syntactic robustness and domain independence. The robustness and domain independence are by-products of a PBP incorporating linguistic principles into the underlying parsing mechanism. That is, a PBP

generates candidate syntactic parses using an extremely general mechanism (X-bar theory). It then uses a number of cooperating principles as constraints on these candidates, thus assuring that the resultant syntactic structures are well formed. One of these constraining principles is the θ-Criterion (Chomsky 1981). The θ-Criterion stipulates that each argument of a verb has a thematic role (i.e., a θ-role) to play in the sentence and that each θ–role of the verb must be assigned. The assignment of these roles is based on the θ–Criterion, in conjunction with the other principles, and does not require any extra rules to be added to the generator.

In contrast to this principled approach, a standard rule-based grammar is created in reaction to sample text and therefore can safely handle only the type of sentences that are in the sample or that are anticipated by its designer. The incorporation of new sentence types requires either additional rules or re-writing of previous rules. The assignment of thematic roles, say for a Case Grammar, requires still more rules. As the syntactic coverage increases, so does the size of the rule-based grammar. At some point the additional rules start interacting with previous rules. A large percentage of the effort then becomes ensuring that new rules do not contradict earlier rules, and that text previously parsed is still parsible. This rule interaction makes a traditional rule-based grammar extremely difficult to make robust, which makes it inappropriate for the current research.

A PBP bypasses this type of rule interaction completely. Each principle in the parser verifies the well-formedness of the sentence in relation to its area of concern. For instance, the Binding Principle is only concerned with sentences that contain anaphors, pronouns or proper nouns. If a sentence contains none of these, then the principle cannot be violated. In such cases the principle considers the parse-structure well formed. If, and only if, none of the principles are violated does the parser consider the whole structure (i.e., sentence) as well formed.

The first step to answering the question then was the development, completion and testing of a partial PBP. A PBP can be developed in modules, one

principle to a module. Each module can be tested separately and, when all the modules are working properly, they can be combined to create the parser.

The principles responsible for generating possible structures are X-bar Theory and the Empty Category Principle (ECP). These two principles acting together generate all possible syntactic structures for a sentence. The vast majority of the structures so generated, however, are ill-formed and will be rejected by another principle later in the parsing process. The most efficacious way to generate the structures is to interleave the generation with the other principles thereby forcing the production of only well-formed structures. This is the approach generally used with PBPs (cf. Fong 1994, Lin 1994). Measuring the impact of a larger lexicon on a complete PBP would be extremely difficult if it used an interleaving approach. Since only well-formed parses are produced, the only measurements available would be the overall speed[1] of the parsing process and the rare, additional, well-formed parse.
This is not rich enough data for our purposes.

We will answer the question then, by measuring intermediate results of the parser while using different lexicons. This will be done by comparing the differences in the total number of possible structures (i.e., by doing exhaustive parsing) produced by an interleaved X-bar/θ-role module using the different lexicons but the same corpus of sentences. This will produce a measurable sense of the impact of the larger lexicon though it is not the ideal way to measure it. It would be preferable to also have the output of just the X-bar module for comparison purposes. The section below on the size of the search space will delineate why this is not possible.

[1] Comparing speeds is especially problematic. The larger lexicon would introduce more syntactic categories which would produce structures that would be rejected by different filters. Since the filters are processed sequentially and not independently, the times would be unrepresentative of the impact of the lexicon.

The lexicon for the first part of the testing is a very small, hand-coded lexicon containing the 155 words found in a small number of sentences (@100). These sentences (Appendix 1) are adapted from the GB literature (compiled by Lasnik and Uriagereka 1988). The sentences provide a "test corpus" against which the modules (X-bar/ECP and θ-filter) are tested. This mimics the standard NLP methodology for accomplishing this type of task (see Gazdar and Mellish 1989). That is, start with a test corpus and insure the developing grammar can cover all the syntactic structures found in the corpus. The difference here is that the grammar being developed is the interaction of the modules (i.e., principles) so what is being tested is the implementation of the principles and not a grammar designed specifically for the test corpus. Second, making the parser larger (by adding more rules) generally increases the coverage. For a PBP however, making the parser larger (by adding more modules) decreases the coverage by filtering out more of the ill-formed structures. Furthermore, the test corpus itself is the result of years of research by numerous people using the Government-Binding approach and therefore illuminates the essence of the principle-based approach. For instance, the pair of sentences below (Lasnik and Uriagereka 1988:31) are part of the corpus that helps to insure that Binding Theory has been properly implemented.

* (a) John believes that Mary likes himself
(b) John believes that Mary likes him

In the first sentence *himself* must have an antecedent (i.e., be bound) in its governing category (the phrase *Mary likes himself*). There is no valid antecedent available and the sentence is ungrammatical. Similarly, in the second sentence, *him* must **not** have an antecedent in its governing category (because it is a pronoun and not reflexive) and there is none available, so the sentence is grammatical. This type of sentence pair is typical for the corpus (See Appendix 1).

We originally thought that it might be possible to acquire a suitable parser for this type of research. We examined a number of PBP parsers (from Dorr, Fong,

Johnson and Lin) but none met all of our requirements. Of the four parsers examined Fong's PAPPI system (Fong 1994) was the most attractive. PAPPI is a robust, multi-lingual PBP written in Prolog. It was not chosen because the structure of its lexicon is quite removed from our own and it is an interleaved parser and therefore would pose problems for the type of testing needed for this research.

As was mentioned earlier, we are using exhaustive parsing for the testing. The only components needed to produce all the possible structures are just the X-bar component and the ECP. A measure of the impact of lexical ambiguity can be obtained by enumerating the number of possible structures generated when the larger lexicon is added and examined after filtering out most of the structures. For reasons covered later, we can not enumerate the structures separately, therefore we are interleaving the X-bar and ECP with one filter. We chose to use the θ-role filter for this purpose because it operates on every well-formed parse (in contrast to bounding theory) and because of the pivotal role that θ-roles play in the rest of our research.

LDOCE

This research uses *Longman's Dictionary of Contemporary English* as the basis for its NLP lexicon. This addresses the issue of domain independence because this MRD contains about 55,000 of the most common words of English, and therefore is appropriate as the base lexicon for any domain. The addition of this lexicon to a complete principle based parser would result in a system capable of analyzing syntax across a variety of domains.

As was mentioned in chapter one, a partial PBP has been written that encompasses a few of the principles. The development of that parser, along with a review of the literature on other PBP systems, has provided evidence concerning the lexical information that is required by a basic PBP. This information was used as a metric against which the contents of LDOCE were compared. The information that

the parser requires that is explicit in LDOCE (syntactic category for example), was noted and the part that is not explicit, was handled in one of two ways. (1) That part that was readily available implicitly (ex., in definitions) was extracted and the method of extraction was noted. (2) The part that was not readily extractable from LDOCE was supplemented from other sources where possible. Sources for this information included Roget's, other MRDs and hand-coding. The results of this comparison are summarized in the next section which presents the various types of information available to the parser, whether the information is required by the parser, the method used to extract the information, and the source from which the information is extracted.

Analysis of LDOCE

Explicit Information

Much of the information required by a full NLP system is explicitly represented in LDOCE. This information includes: the pronunciation; whether nouns have plural forms; level of usage (formal, informal, slang); country in which the word is used (America, Britain, Australia); and definitions. The parser requires only a small part of this information, namely:

Word and syntactic category, which is required by the X-bar component. This includes the normal categories (i.e., noun, verb) as well as marking of the type for pronouns (i.e., nominative, reflexive, possessive).

Subcategorization. Required by the X-bar component and the projection principle. Part of this information is represented explicitly in the LDOCE grammar codes and is fairly diverse: verb transitivity, clausal objects (as opposed to nominal objects), attributive adjectives, and verbs that cannot occur in progressive form (ex., not, "I am hating football.").

Implicit Information

Case roles. Required by the case filter. Rather than derive the case roles from the subcategorization information, the roles were taken directly from rules given in Sells (1985).

θ-roles. Required by the θ-criterion. The θ-roles were derived by a computational analysis of the LDOCE definitions. The description of this analysis is beyond the scope of the current work.

Other Sources

Roget entry number. While not specifically required by the parser, the mapping of LDOCE definitions to Roget's hierarchy provides a richer semantics that would be usable by a complete NLP system.

It was stated above that the information in LDOCE was converted to the form needed by the parser. This task is far from trivial, and should not be underestimated. Guo (1989) wrote his dissertation on the conversion of an earlier version of LDOCE, and using his dissertation greatly aided in solving some of the problems of organization and processing. Much of this processing was previously accomplished by Woojin Paik for the DR-LINK project (Liddy and Paik 1991). The conversion from the original printer's tape to ASCII files and the simplification of many print-dependent codes were accomplished for that project. The output from that conversion was subsequently reformatted into a Prolog-like database and a sample lexicon was created as part of the feasibility study for this proposal. The form of the database has not subsequently changed significantly.

Implementation of PBP

The parser that was written for the feasibility study was updated to answer this research question. As was presented in chapter 2, the parser uses two main grammar rules.[2]

XP → specifier XB modifier
XB → preadjunct X argument

These rules have to be considered as a high level abstraction of what the parser actually looks like but the parser is not as removed from these rules as we might wish. The parser, while being more sophisticated than the presentation here indicates, is, nonetheless, simple to the point of naïveté. It does, however, serve its purpose for this question. The parser is written in Prolog and uses Definite Clause Grammar (DCG) and Prolog's top-down, depth-first, left-to-right control mechanism. There is no penalty for using this mechanism in that we are doing an exhaustive search through the parsing space. Other mechanisms were considered (including a variation of Johnson's bottom-up chart parser) but in the end simplicity won over sophistication.

As it is, the parser is quite respectable in speed. It produces candidate parse trees on the average of about 100 per second while running in ALS Prolog under DOS on a 486/66 PC. While speed was not a consideration in the design of the parser it probably was a by-product of the parser being designed to do exhaustive parsing. A "walk-through" the parsing process should clarify this.

[2] These two rules are phrase structure rules and therefore our grammar could be labeled as a phrase structure grammar, as proponents of lexicalized TAGs (Vijay-Shanker and Joshi 1985) might be quick to point out. As mentioned earlier, GB is a very impoverished phrase structure grammar. We feel that generally characterizing it as a phrase structure grammar would be akin to calling a collection of two books a library – not incorrect but a bit excessive.

Parsing Process

The process starts by reading in a sentence and converting the sentence to a list. Each word in the list is then assigned all the information in the lexicon associated with its most likely sense (i.e., its first occurrence in LDOCE). The parser then tries to instantiate the simplest version of the X-bar template possible for the sentence. That is, it starts by assuming that there are no specifiers, modifiers or preadjuncts and only those arguments subcategorized for (the subcategorization ensures that a θ-role can be assigned). If this succeeds, the parse is output. The parser then "backtracks" to its last choice point and tries to re-do the choice perhaps by allowing a modifier or preadjunct. This continues until no further candidate trees are possible with the current word senses. The parser then tries to use empty categories for the arguments using a licensing scheme based on Johnson and Kay (1993). Finally, after all possible candidate trees using licensed empty categories are output, the parser retries using a different word sense for each word.

Size of the Search Space

The size of the output can now be perceived. There are three contributing factors: (1) the number of senses for each word, (2) the number of slots in the X-bar template, and (3) the number of empty categories per sentence.

The number of senses for each word is a factor in that each combination of syntactic category produces a different tree. That is, if a two word sentence were being processed (ex., *pat dreams*) and each word had two senses ([a,b] and [c,d] respectively) then there would be four possible combinations to consider: [a, c], [a, d], [b, c], and [b, d][3]. If each word had three senses then there would be nine possible combinations. However, the number of possible trees is much greater than that because each word can occupy any of three of the five slots in the X-bar

[3] It is slightly worse than stated. Two senses can share the same syntactic category yet require different trees because of differences in their subcategorization or θ-role frames. This affect, though, is much smaller than the combinatoric effect.

template. (The X-bar template is reshown as Figure 4.2). For example, in a two word sentence, the first word can be a specifier or preadjunct to the second word (or the head) and the second word can be the head, or an argument or modifier to the first word. This results in five slot arrangements, each of which can take any of the combinations produced by the word senses. For the two word example sentence used above that would result in 20 (4x5) possible trees. This can be shown by numbering the slots 1-5 such that specifier=1, preadjunct=2, X=3, argument=4 and modifier=5. Given that, the list [b3, d5] would represent the tree having **b** as the head and **d** a modifier.

The ten possible trees for sense **a** would then be given by

1. [a1, c3]
2. [a1, d3]
3. [a2, c3]
4. [a2, d3]
5. [a3, c4]
6. [a3, d4]
7. [a3, c5]
8. [a3, d5]
9. [a3, c3]
10. [a3, d3]

b would have the same number thereby comprising the 20 total combinations.

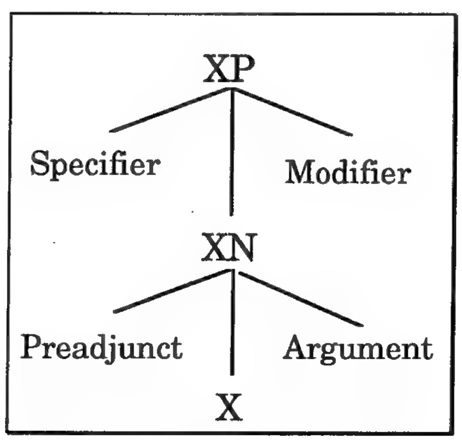

Figure 4.2. The X-bar Template

Once again, though, the number of possible trees is actually greater than this because we allow empty categories. Remember that an empty category represents an elided or moved constituent so that any of the five slots can be filled with an empty category. Furthermore, there can be multiple empty categories per sentence. If we assume that there are no more empty categories than there are words in the sentence (a generally safe assumption) there are still an extremely large number of candidate parse trees produced. For each of the 20 trees produced above there would be six trees produced using one empty category. For the [a1, c3] case these would be

1. [a1, e2, c3]
2. [a1, c3, e4]
3. [a1, c3, e5]
4. [a1, c3, e3]
5. [e3, a1, c3]
6. [a1, c2, e3]

Where **e** is an empty category. Note that the categories do not have to be in the same tree (there is always more than one tree per sentence). Allowing two empty categories produces four trees for each of the trees produced with one empty category. For the [a1, e2, c3] example,

1. [a1, e2, c3, e4]
2. [a1, e2, c3, e5]
3. [a1, e2, c3, e3]
4. [e3, a1, e2, c3]

The simple two-word, two-senses per word sentence, then, has at least 480 (4x5x6x4) possible trees. In general, the number of possible trees is greater than the square of the product of the number of syntactic categories of the words in the sentence. For a ten-word, two-senses per word sentence the number of possible trees would be over 1,000,000 (2^{20}). For any normal sentence, generating all possible trees producible by X-bar/ECP would be computationally intractable. It is for this reason that we integrated the θ-criterion with the X-bar module. The

integration simply involves limiting argument structures only to verbs that subcategorize for an argument. This restriction assures that there is a θ-role to assign and reduces the output to the same size as that which would be produced by an unintegrated filter.

Results

A comparison of the sizes of the two lexicons is given in Table 4.1 and the results of the testing are in Table 4.2. The LDOCE lexicon has only 1.6 times as many word senses as the hand coded one which seems a minimal increase. However, the average number of combinations of syntactic categories (including empty categories but disregarding sentence structure) produced by LDOCE is just under 17 times that of the smaller lexicon. The difference in the average number of structures produced by the X-bar/θ-role module concomitantly increased from 2,738 to 47,875, or just over 17 times.

	Hand-coded	***LDOCE***	**x Increase**
Number of Words	151	151	1
Number of Senses	165	266	1.61
Average Number of Senses	1.09272	1.76159	1.61

Table 4.1. Size of the Two Lexicons

	Hand-coded	***LDOCE***	**x Increase**
Average Number of Structures	2738.725	47875.522	17.48
Minimum	4	5	1.25
Maximum	82719	2953705	35.707
Median	249	1366	5.486
Average Number of Combinations	1832	30062	16.40

Table 4.2. Number of Structures Produced

At first glance, these results are not exceedingly encouraging. The average number of structures produced by the parser is roughly equal to the average

number of combinations of syntactic categories for the sentences. This probably means that any increase in either the length of the sentence or in the number of senses per word would have an exponential effect on the parser.

The actual effect is quite a bit smaller than that. Figure 4.3 shows the one sentence (*This article is too illogical to read without laughing at*) that is skewing the average much higher than it should be. If the median is considered, rather than the average, then the number of structures produced

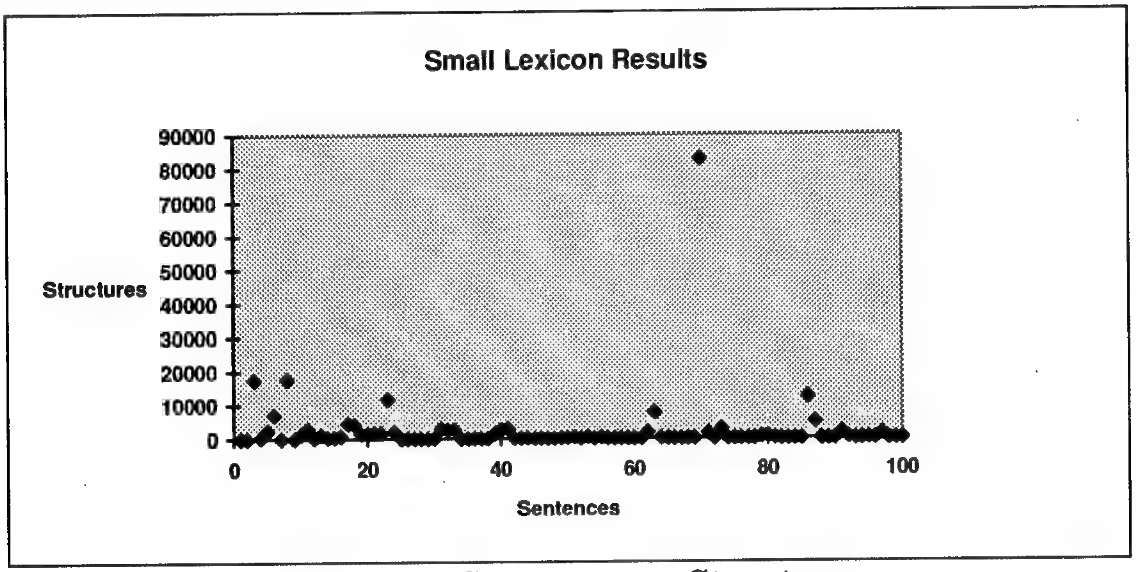

Figure 4.3. Sentences vs. Structures

using LDOCE is much smaller than the average number of combinations required. In fact, it is even smaller than the average number of combinations required for the smaller lexicon, and this using exhaustive parsing and only one filter. A normal PBP would do neither. The 1000 or so structures produced are a result of the exhaustive nature of this particular PBP. Normally, only those structures that are completely acceptable as sentences would be produced. If the results obtained for the θ-filter hold for the other filters then the effect of the larger lexicon would be minimal for most sentences.

Summary

The methodology used is a systems approach. This type of approach has been used effectively by numerous researchers in NLP, IR and related fields. The overall approach is quite similar to those traditionally used in IR research while being somewhat more test intensive than traditional NLP research. The main assumption of the methodology is that building the systems, designing and describing the algorithms, testing and analyzing the results, is a rigorous approach to answering the questions in a way that is both useful and repeatable.

This chapter dealt with the impact of a larger lexicon on a principle based parser. To measure the impact, a partial PBP was constructed consisting of the X-bar/ECP structure generator and one filter (θ-criterion). The parser was then exhaustively run with two different lexicons while keeping the corpus constant. The results indicate that an increase in the size of the lexicon (i.e., number of senses per word) produces a substantial, though still tractable, increase in the number of structures accepted. These results are encouraging yet are not as exhaustive as we might wish.

There are a number of ways the testing could be extended. More filters, or different filters, could be added to the parser. This would ensure that the results were not a quirk of the θ-criterion. Our previous work with the case filter, though, leads us to believe that the case filter would have very similar results to the θ-filter.

A different lexicon could be used. This could provide even more senses/word than LDOCE did thus providing more data for a trend analysis of the impact of the lexicons. We have not done any comparative testing of machine readable dictionaries to ensure that LDOCE is typical as to the average number of senses per word.

A different corpus (or corpora) could be used. The corpus selected was chosen because it completely exercises X-bar theory. It is not typical of any domain and

specifically it is not typical of corpora designed to exercise IR techniques. IR corpora might be problematic to a PBP especially in respect to the presence of much longer sentences, multiple conjunctions and compound nominals. The testing done within this research is only a first step toward that needed to be done to demonstrate the generalizability of the results to typical IR domains and tasks. Again, due to limitations of time and resources, we have not tested other corpora.

It was our intent to show that a machine readable dictionary has sufficient information to perform as the lexicon for a principle based parser and that using the derived lexicon does not unduly affect parser performance. We believe that the results support those claims.

Chapter 5

Discussion and Conclusions

We are not certain, we are never certain. If we were we could reach some conclusions, and we could, at last, make others take us seriously.

Albert Camus

This research has investigated components of natural language processing (NLP) that should eventually provide a system with the ability to classify and categorize concepts and relations in a more domain independent manner than is currently possible. One goal of such a system is the improvement of the ability of information retrieval (IR) to provide relevant information and texts to users. Whether the NLP components investigated here actually provide IR with these capabilities remains to be seen. The investigation of these components has, nonetheless, provided us with a number of insights, techniques and results that can be exploited by researchers in NLP, IR and other fields.

Impact of the Lexicon

We have examined the combination of a broad coverage, domain independent parser with a lexicon derived from machine readable resources. The parser, a principle based parser, is of a class normally associated with linguistic studies and not with fielded NLP systems. Fong (1994) has demonstrated that a PBP can be robust if used with a controlled vocabulary. Lin (1994) has shown that a PBP can use a large lexicon if the structures the parser is allowed to produce are severely limited. We have shown that the parser remains computationally tractable with a large lexicon even allowing a more relaxed control over the structures.

These results may encourage more IR/NLP researchers to use a PBP especially when supported by the intrinsically attractive features of the approach. One of the problems facing NLP researchers that have used parsing for IR is the need for the parser to produce some syntactic structure regardless of the type of text encountered. This has resulted in some ingenious relaxation rules and other methods to generalize the rule base to handle sentence fragments, ill-formed input and the like. It was shown in Chapter 4 that the X-bar/ECP module will generate myriad structures for every sentence thus the relaxation technique needed to

handle ill-formed input is simply to by-pass one or more filters. The subsequent exploitation of those structures by IR remains an interesting question.

Future Research

The work, as presented in Chapter 4, is just the first step toward the total analysis of the impact of a larger lexicon on the parser. While LDOCE is imminently richer than our hand-coded lexicon – yielding lots of lexical ambiguity, LDOCE is an abridged (i.e., not unabridged) dictionary. If the same tests could be re-run using data from an unabridged dictionary, and thus all the known syntactic categories for each word, then our claim of a larger lexicon causing minimal impact could be stated with more conviction.

For the same reason, the tests should be re-run using more and different filters. We chose the θ-Criterion because of its special relevance to the rest of our work. It has no special significance beyond that. The rest of the filters should be tested both separately and together. This would again add credence to any claims made.

A wider range of text would also help with claims of generalizability. We have assumed that the text tested was at least as difficult for the parser as most text likely to be encountered. Testing it with scientific abstracts or other text associated with common IR tasks would be of benefit.

Of course, to be really general, the above tests would have to be redone with a language other than English. The Government Binding literature has shown the benefit of such testing and the benefits should hold for a PBP.

Overall Impact of the Research

The System and IR

The impact of a machine readable dictionary on a principle based parser has been examined. This combination is just on component of a design for an NLP system with a domain independent parser, a solid basic lexicon and extended semantics which we hope will eventually prove useful for information retrieval. The relation of that system to IR really depends on how IR is defined. Let us examine the relation for three aspects of IR: document retrieval, conceptual information retrieval and lexical browsing.

Document Retrieval

If document retrieval is kept separate from conceptual retrieval then our system would have a somewhat limited role to play. The retrieval of relevant documents from a database is usually done by using words or phrases likely to occur in the relevant documents but not in non-relevant ones. In such cases using proximity matching (i.e., words occurring within 3 or 4 words of another) produces greater recall than exact string matching.[1] A reason for this can be seen by looking at a very simple case. If one were looking for documents on *lexical browsing* then relevant documents might have the following phrases.

lexical browsing
browsing of lexical materials
browsing of materials, especially lexical ones

This tendency of natural language to rephrase to increase interest and comprehension makes exact string matching difficult.

The apparent use of the system would probably not be to increase recall but rather to increase precision. The system, by "understanding" the sentences that the

[1] Metzler and Haas (1989).

key words or phrases occur in, might be able to filter out those documents that are really extraneous to the search.[2] Such a demonstration for general IR is obviously beyond the scope of the current research but is well worth pursuing.

Conceptual Information Retrieval

The ultimate use of our system in conceptual information retrieval seems more likely. Myaeng et al. (1994) conducted experiments within a Conceptual Graph (Sowa 1984) framework using extended thematic roles and concept-relation-concept triples. They felt that the mixed retrieval results were largely due to errors propagated through the stages of text processing. Errors which they thought a full parser would be less prone to make. While a traditional phrase structure grammar could be written to handle the text they were dealing with, general IR would require more; a robust, domain independent parser coupled with a large domain independent lexicon. We have shown why a PBP fills the requirement for the parser. The lexicon is another matter.

LDOCE has 55,000 or so words of everyday English, Roget's includes up to 250,000 more words and phrases. This, one might argue, proves domain independence. It does and it doesn't. The system lexicon is independent of domain in that the words in the lexicon are apt to occur in any domain with little or no changes in their meaning. Words often drift semantically between domains (i.e., acquire new meanings or lose old ones) but seldom are they completely redefined. For the lexicon to be completely useful in all domains though, it would have to contain, or be able to acquire, all the words in each domain. In this sense our lexicon is not domain independent. A vast majority of these domain specific words are nouns (Miller 1991) and thus are not directly affected by thematic roles (though they would be by selectional restrictions). The LDOCE/Roget lexicon then provides

[2] This is similar to the approach used by Jacobs and Rau (1988, 1990) for a smaller domain.

not an ultimate domain independent lexicon but a solid foundation for any domain specific lexicon required by a conceptual retrieval system.

The semantics presented are similar to that needed for conceptual information retrieval (Chaffin and Herrmann 1987, Sabah and Vilnat 1991). Certainly the ability to abstract relations at different levels is an important one for this task. Whether this type of semantics is the type needed for general IR remains to be seen but semantics of this type have proven useful in smaller domains[3]. One possibility is that this is the type of semantics needed for general IR but Roget's may not be the tool to provide it. Again, further research is warranted.

Lexical Browsing

Along with conceptual retrieval, lexical browsing seems an appropriate application for the system as presented. Longman's Lexicon (Longman 1981) is the integration of a dictionary with a thesaurus so it appears that an integrated system is required to construct a browser (or it must be constructed by hand). But what is the relationship of a browser to IR?

A browser can be seen as another form of IR. It is not one in which a user is looking for relevant documents but it is one where a user is looking for relevant information. The utility of a browser to a newcomer in a domain would be helpful in the context of learning a new language. The same utility might allow even experienced personnel to find the information they need quickly and easily in other systems. Consider an example from an air force domain.

A browser could provide a person new to the domain an easy way of becoming acquainted with new terms and relationships. For example, the relationship between pilot and navigator, navigator and radar, radar and weather could all be readily and easily explored. The same possibility is also useful to more experienced

[3] Myaeng and Liddy (1993) showed improved results using semantic relations over systems that used NLP techniques without them.

personnel. If a radar operator encounters a new (to them) interference pattern on their scope they could use a browser to check the relationship of "running rabbits" (a type of interference pattern) to weather and to electronic jamming. Such a tool, if properly designed, would provide not only on the job training but would also become a useful problem solving tool by providing highly germane information easily and readily. The construction of such a tool would be greatly aided by the system as presented in this research.

Browsing is quickly becoming a standard operating procedure for those people "surfing the net" and especially for those on the World Wide Web. For any of those that have done a net search for information, it is probably apparent that the problems of precision and recall have not been solved.

The Research and Testing

This research has presented components of an NLP system that extend the domain independence of a principle based parser with two levels of semantics. During the investigation of these components we made some interesting discoveries, reinforced earlier findings, and verified some assumptions. Most of the ways in which we would like to have this research impact others have already been discussed: an increase in use of principle based parsing; a closer relationship between theoretical linguistics and natural language processing; as a step toward conceptual information retrieval; and various influences on lexicography. These points will not be belabored here. There is one further point that we would like to discuss; the methodology used in this research.

This research was not intended to develop new methodologies for NLP, linguistics, IR or any other field. Yet the methodology used, that of rigorous testing, is almost foreign to NLP researchers and even to the more general artificial intelligence (AI) researcher community. This sort of testing is not something

developed just for the present research. In fact, it is *de rigeur* in many fields, including IR. It is not even something new as applied to NLP and AI. Cohen and others (Cohen and Howe 1988, Neal et al. 1991) have been striving to develop a more rigorous methodology for AI but little progress has been made in convincing others of the need. Perhaps the current research can demonstrate that need.

Most of the interesting discoveries made during the course of this research are a direct result of the in-depth testing that was used. The need to integrate the θ-filter with the parser was uncovered because we were trying to test the impact of the lexicon by keeping the parser invariant while varying the lexicon.

Of all the lessons learned from this research we find this to be one of the most valuable – NLP research needs more rigorous testing. Not just so we can have repeatable results but so that we can discover all the little bits of knowledge that we have been missing and maximize the potential of our research.

Appendix 1 - GB test corpus
(adapted from Lasnik and Uriagereka 1988)

After visiting Jean who did you hire.
Everyone that Tim knows he likes.
I am eager for Kim to be here.
I am proud of Kim.
I am proud that Kim is here.
I believe Kim is here.
I believe Kim to be a pathological liar.
I believe Kim to be here.
I believe Kim to be intelligent.
I believe sincerely Kim to be here.
I persuaded Kim to leave.
I sincerely believe Kim.
I tried Kim to win.
I tried to leave.
I tried to understand the problem.
I tried to win.
I want Kim to win.
I want to be clever.
I want to visit you.
I want to win.
I wanted it to rain.
I wanted Kim to leave.
I wanted the bus to arrive on time.
I wonder who you think Kim said you will see.
I wonder who you will see.
Is raining.
It is illegal for Kim to park here.
It is important for somebody or other to understand the problem.
It is important to be here.
It is important to eat.
It is likely Kim to be here.
It is likely that Kim is here.
It is likely that Kim will win.
It is raining.
Jean will solve which problem.
Jean wonders who saw Kim.
Jean wonders who saw what.
Kim hit Jean.
Kim is certain to see this.

Kim is likely to park here.
Kim is likely to win.
Kim is too dumb to talk to.
Kim likes everyone.
Kim often arrives late.
Kim seems to be crazy.
Kim slept.
Kim thinks he likes Jean.
Kim thinks he washed himself.
Kim told Jean that they should leave.
Kim was arrested after leading the demonstration.
Kim was arrested by the police.
Kim was arrested.
Kim was believed to be clever.
Kim was believed to be intelligent.
Kim was persuaded to leave.
Kim's destruction of the building.
Kim's teacher can not stand Kim.
Kim's teacher can not stand the oaf.
Running away upset Jean.
Someone likes everyone.
The problem was solved.
The teacher fell on the floor after reading the book.
The teachers think that pictures of each other will be on sale.
Their pictures of each other are nice.
They arrested Kim.
They like each other.
They like themselves.
They read each other's books.
They read their books.
They think that it be likely that pictures of each other be on sale.
They want to visit each other.
This article is too illogical to read without laughing at.
Tim believes Mary to like him.
Tim believes that Mary likes herself.
Tim believes that Mary likes him.
Tim likes him.
Tim likes himself.
Tim likes Mary's pictures of him.
Tim likes pictures of himself.
Tim saw pictures of himself.

We think that I will win.
What was filed without being read.
What will you read.
Which problem will Jean solve.
Which report did you file without reading.
Who did you give a picture of Jean to.
Who did you give a picture of to Jean.
Who did you mention that Jean believes that you saw.
Who do you think saw Jean.
Who do you think that Kim saw.
Who do you want to win the race.
Who left after you insulted him.
Who left.
Who resigned before we could fire him.
Who that Tim knows does he like.
Who thinks Mary likes him.
Who tried to win the race.
Whose mother does Kim like.
Why do you think Kim left.
Why don't you know which worker Kim fired.

Bibliography

Abney, S. and J. Cole (1985) "A Government-Binding Parser", *Proceedings of N.E.L.S. 16.*

Ahlswede, T.E. (1988) *Syntactic and Semantic Analysis of Definitions in a Machine-Readable Dictionary*, Ph.D. Dissertation, Illinois Institute of Technology.

Akkerman, E. (1989) "An independent analysis of the LDOCE grammar coding system," In Boguraev and Briscoe (1989).

Anderson, J.M. (1971) *The Grammar of Case: Towards a Localistic Theory*, Cambridge University Press: Cambridge, U.K.

Barton, G.E. , R.C. Berwick and E.S. Ristad (1987) *Computational Complexity and Natural Language*, The MIT Press: Cambridge, MA.

Becker, D. (1981) "Automatic Language Processing" *ARIST* 16.

Berrut, C. (1990) "Indexing Medical Reports: The RIME Approach", *Information Processing and Management*, 26(1).

Berrut, C. and P. Palmer (1986) "Solving grammatical ambiguities within a surface syntactical parser for automatic indexing." In F. Rabitti (ed.), *Proceedings of the Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Association of Computing Machinery, New York: 123-130.

Berwick, R. C. and S. Fong (1990). "Principle Based Parsing: Natural Language Processing for the 1990s". In P.H. Winston and S.A. Shellard (Eds.), *Artificial Intelligence at MIT: Expanding Frontiers.* pp. 286 - 325. The MIT Press: Cambridge, MA.

Berwick, R. and A. Weinberg (1984) *The Grammatical Basis of Linguistic Performance.* MIT Press: Cambridge, MA.

Blair, D. (1990). *Language and Representation in Information Retrieval.* Amsterdam: Elsevier Science Publishers.

Boguraev, B. and T. Briscoe (1989) *Computational Lexicography for Natural Language Processing*, Boguraev, B. and T. Briscoe (eds.), Longman: New York.

Braun, S. and C. Schwind (1976) "Automatic, Semantics-Based Indexing of Natural Language Texts for Information Retrieval Systems", *Information Processing and Management*, Vol. 12, 147-153.

Buckley, C. (1985) *Implementation of the SMART Information Retrieval System.* Technical Report TR85-686, Department of Computer Science, Cornell University: Ithaca, New York.

Byrd, R., N. Calzolari, M. Chodorow, J. Klavans, M. Neff, and O. Rizk (1987) *Tools and methods for computational lexicology*, RC-12642, IBM: Yorktown Heights, NY.

Carroll, J. and C. Grover (1989) "The derivation of a large computational lexicon for English from LDOCE," In Boguraev and Briscoe (1989).

Chafe, W.L. (1970) *Meaning and the Structure of Language*, Chicago University Press: Chicago, IL.

Chaffin, R. and D.J. Herrmann (1987) *Relation Element Theory: A New Account of the Representation and Processing of Semantic Relations*, in D. Gorfein and R. Hoffman, (Eds.), Memory and Learning: The Ebbinghaus Centennial Conference, Lawrence Erlbaum, Hillsdale, NJ.

Chomsky, N. (1957) *Syntactic Structures,* Mouton: The Hague.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*, MIT Press: Cambridge, MA.

Chomsky, N. (1981) *Lectures on Government and Binding: The Pisa Lectures.* Foris Publications: Amsterdam.

Chomsky, N. (1982) *Some Concepts and Consequences of the Theory of Government and Binding.* MIT Press: Cambridge, MA.

Chomsky, N. (1986) *Knowledge of Language: Its Origins, Nature, and Use,* Praeger Publishers, NY.

Chomsky, N. (1986) *Barriers.* MIT Press: Cambridge, MA.

Cohen, P.R. and A.E. Howe (1988) *Towards AI Research and Methodology: Three Case Studies in Evaluation.* COINS Technical Report 88-31, University of Massachusetts, Amherst, MA.

Collins English Dictionary (1979) *Collins English Dictionary on CD-ROM.* William Collins Sons & Co., Ltd. London.

Cook, W. (1989) *Case Grammar Theory.* Georgetown University Press: Washington.

Correa, N. (1988) *Syntactic Analysis of English with Respect to Government-Binding Grammar*, Ph.D. Dissertation, Syracuse University: Syracuse, New York.

Croft, B. and D. Lewis (1987) "An Approach to Natural Language Processing for Document Retrieval" In C.T. Yu and C.J. van Rijsbergen, (Eds.), *Proceedings of the Tenth Annual International ACM SIGIR Conference.* Association of Computing Machinery, New York: 26-32.

Cruse, D.A. (1986) *Lexical Semantics*, Cambridge University Press: Cambridge, UK.

Cullingford, R. (1986) *Conceptual Dependency.* Addison-Wesley: New York.

De Fude, B. (1986) *Etude et réalization d'un système intelligent de recherche d'informations: le prototype IOTA.* Ph.D. thesis, Grenoble University, France.

de Saussure, F. (1901) *Cours de linguistique générale*, Bally, C., A. Sechegaye and A. Riedlinger (eds.), W. Baskin (trans.), McGraw-Hill: New York, 1959.

Dick, J.P. (1991) *On the usefulness of conceptual graphs in representing knowledge for intelligent retrieval.* Proceedings of the Sixth Annual Workshop on Conceptual Graphs. July 1991, Binghamton, NY.

Dillon, M. and A. S. Gray (1983) "FASIT: a Fully Automatic Syntactically based Indexing System." *Journal of the ASIS*, 34(2), pp. 99-108.

Dorr, B. (1987) *UNITRAN: A Principle-Based Approach to Machine Translation*, AI Technical Report 1000, Master of Science, Department of Electrical Engineering and Computer Science, MIT: Cambridge, MA.

Dowty, D., R. Wall and S. Peters (1981) *Introduction to Montague Semantics.* D. Reidel Publishing: Dordrecht, the Netherlands.

Fagan, J. L. (1987) "Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-syntactic Methods." In C.T. Yu and C.J. van Rijsbergen, (Eds.), *Proceedings of the Tenth Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval. Association of Computing Machinery, New York: 91-101.

Fagan, J. L. (1988) *Experiments in Automatic Phrase Indexing for Document Retrieval: a Comparison of Syntactic and Nonsyntactic Methods.* PhD Dissertation, Cornell University, Ithaca, New York.

Fillmore, C. (1968) "The Case for Case". In *Universals in Linguistic Theory*, Bach, E. and R. Harms (Eds.), Holt, Rinehart and Winston: New York.

Fillmore, C. (1977) "The Case for Case Reopened". In *Syntax and Semantics*, Cole, P. and J. Sadock (Eds.), Academic Press: New York.

Fong, S. (1989) *Principle-based Parsing and Principle-Ordering.* Report AIM-1156, Artificial Intelligence Laboratory, MIT, Cambridge, MA.

Fong, S. (1991) "The Computational Implementation of Principle-Based Parsers". In Berwick, R. C. et al (eds) *Principle-Based Parsing: Computation and Psycholinguistics*, Kluwer Academic Publishers, Hingham, MA.

Fong, S. (1994) The PAPPI system. Available from sandiway@research.nj. nec.com

Gazdar, G., A. Franz, K. Osborne and R. Evans (1987) *Natural Language Processing in the 1980s*, Center for the Study of Language and Information, Stanford, CA.

Gazdar, G. and C. Mellish (1989) *Natural Language Processing in Prolog.* Addison-Wesley: New York.

Grosz, B. and C. Sidner (1986) "Attention, intentions and the structure of discourse," *Computational Linguistics*, 12(3), 175-204.

Gruber, J.S. (1965) *Studies in Lexical Relations*, Ph.D. Dissertation, MIT: Cambridge, MA.

Guo, C-M. (1989) *Constructing a Machine-Readable Dictionary from "Longman Dictionary of Contemporary English"*, Ph.D. Dissertation, New Mexico State University: Las Cruces, NM.

Haegerman, L. (1991) *Introduction to Government and Binding Theory*. Blackwell, Oxford. Second Edition 1994.

Heidorn, G.E. (1972) *Natural Language Inputs to a Simulation Programming System*, Technical Report NPS-55HD72101A, Naval Postgraduate School: Monterey, CA.

Heidorn, G.E. (1975) "Augmented Phrase Structure Grammar" in Schank, R.C. and B.L. Nash-Webber (eds.), *Theoretical Issues in Natural Language Processing*. Association of Computational Linguistics.

Hirst, G. (1987) *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press: Cambridge, UK.

Jackendoff, R. (1972) *Semantic Interpretation in Generative Grammar*. MIT Press: Cambridge, MA.

Jackendoff, R. (1990) *Semantic Structures*. MIT Press: Cambridge, MA.

Jacobs, P. S. and L. F. Rau (1988) "Natural Language Techniques for Intelligent Information Retrieval" *Proceedings of the 1988 ACM SIGIR Conference on Research and Development in Information Retrieval*. Association of Computing Machinery, New York: 85-99.

Jacobs, P. S. and L. F. Rau (1990) "SCISOR: Extracting Information from On-line News," *Communications of the ACM*, Vol. 33, No. 11, November 1990.

Jensen, K. (1986) *PEG 1986: A Broad-Coverage Computational Syntax of English*. Research Report, IBM Thomas J. Watson Research Center, Yorktown Heights, New York.

Johnson, M. and M. Kay (1993) “Parsing and Empty Nodes.” In C.G. Brown and G. Koch (eds.), *Natural Language and Logic Programming III*, North Holland, Amsterdam.

Kittredge, R. and J. Lehrberger (1982) *Sublanguages: Studies of Language in Restrictive Domains*. Walter DeGruyter: New York.

Kuhns, R. J. (1990) "Automatic Indexing and Government-Binding Theory". *Proceedings of COLING 90*.

Lasnik, H. and J. Uriagereka (1988) *A Course in GB Syntax*, MIT Press: Cambridge, MA.

Lau Tzu (550 B.C.) *Tao Te Ching*, trans. Gia-Fu Feng and J. English, Vintage Books: New York, 1972.

Liddy, E. D. and D. M. Lauterbach (1993) *Automatic Construction of a Semantic Lexicon for use in Natural Language Processing Systems*. Final Report for Research Initiation Program. Rome Laboratory, Griffiss AFB, NY.

Liddy, E. D. and Paik, W. (1991) "Automatic Recognition of Semantic Relations in Text". In *Proceedings of the Informatics II Conference*, York, England.

LDOCE (1987) *Longman Dictionary of Contemporary English.* Longman: Harlow, UK.

Lin, D. (1994) "PRINCIPAR – An Efficient, Broad-coverage, Principle-based Parser." In Proceedings of Coling 94.

Longman (1981) *Longman Lexicon of Contemporary English.* Longman: Harlow, UK.

Lu, X. (1990) *An Application of Case Relations to Document Retrieval.* Ph.D. Dissertation, University of Western Ontario, Canada.

McCawley (1988) *The Syntactic Phenomena of English.* University of Chicago Press, Chicago, IL.

Mc Hale, M. L. (1991) *The Production of a Parser for Longman's Dictionary of Contemporary English.* Presented at the IEEE Dual-Use Technology Conference, SUNY Utica-Rome, Utica, NY.

Mc Hale, M. L. (1991a) Natural Language Processing and Information Retrieval. RL-TM-91-29. Rome Laboratory, Griffiss AFB, New York.

Mc Hale, M. L. (1995) *Combining Machine-Readable Lexical Resources with a Principle-Based Parser*, Ph.D. Dissertation, Syracuse University, NY.

Metzler, D. and S. W. Haas (1989) "The Constituent Object Parser: Syntactic Structure Matching for Information Retrieval", *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, Boston, MA.

Metzler, D. and S. W. Haas (1990) "Conjunction, Ellipses and Other Discontinuous Constituents in the Constituent Object Parser", *Information Processing and Management*, 26(1).

Miller, G.A. (1991) *The Science of Words*, Scientific American Library: New York.

Montague, R. (1972) "The Proper Treatment of Quantification in Ordinary English". In **Hintikka, J. et al** (eds.) *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics* Dordrecht: Reidel pp 221-242

Montague, R. (1974) "English as a Formal Language". In R. Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press: New Haven, CT.

Montgomery, C. A. "Linguistics and Information Science". *Journal of the American Society for Information Science* 23(3): 195-219.

Morris, J. and G. Hirst (1991) "Lexical Cohesion by Thesaural Relations as an Indicator of the Structure of Text," *Computational Linguistics*, 17(1).

Myaeng, S. H., C. Khoo and M. Li (1994) "Linguistic Processing of Text for a Large-Scale Conceptual Information Retrieval System". In W.M. Tepfenhart, J.P. Dick and J.F. Sowa (eds). *Conceptual Structures: Current Practices*. Proceedings of the Second International Conference on Conceptual Structures, College Park, MD. Lecture Notes in Artificial Intelligence, 831. Springer-Verlag, Berlin.

Myaeng, S. H. and E. D. Liddy (1993) "Information Retrieval with Semantic Representation of Texts." In *Proceedings of the 2nd Symposium on Document Analysis and Information Retrieval*, Las Vegas, April 26-28, pp. 201-215.

Neal, J.G., Feit, E.L. and C.A. Montgomery (1991) "Benchmark Investigation/Identification Project: Phase I." In J.G. Neal and S.M. Walter (eds.), *Natural Language Processing Systems Evaluation Workshop*. RL-TR-91-362. Rome Laboratory, Griffiss AFB, NY.

Oxford English Dictionary (1992) *Oxford English Dictionary on CD-ROM*. Oxford University Press, Walton Street, Oxford.

Papegaaij, B. C. (1986) *Word Expert Semantics: an Interlingual Knowledge-Based Approach*. V. Sadler and A.P.M. Witkam (eds.) Foris Publications: Dordrecht, Netherlands.

Pereira, F.C.N. and D.H.D. Warren (1980), "Definite Clause Grammars for Language Analysis - a Survey of the Formalism and a Comparison with Augmented Transition Networks", *Artificial Intelligence*, 13 231-278.

Pereira, F.C.N. (1982) *Logic for Natural Language Analysis*, Ph.D. Dissertation, Edinburgh University, Edinburgh, Scotland.

Pustejovsky, J. and B. Boguraev (1991) "Lexical Knowledge Representation and Natural Language Processing," *IBM Journal of Research and Development*, 35(4).

Random House Dictionary (1980) *The Random House College Dictionary, Revised Edition*. Random House, Inc. New York.

Rieger, C. and S. Small (1968) "Word Expert Semantics." *Proceedings of IJCAI 1968.*

Roget, P.M. (1852) *Thesaurus of English Words and Phrases, Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition.*

Roget (1977) *Roget's International Thesaurus, Fourth Edition*. R.L. Chapman (ed.), Harper & Row: New York.

Roget (1992) *Roget's International Thesaurus, Fifth Edition*. R.L. Chapman (ed.), Harper Collins: New York.

Sabah, G. and A. Vilnat (1991) "Flexible Case Structure Implemented into a Deterministic Parser". in *Proceedings of the Sixth Annual Workshop on Conceptual Graphs*: Binghamton, NY.

Sager, N. (1981) *Natural Language Information Processing*, Addison-Wesley: New York.

Salton, G. (1991) Presentation at the NLP-IR Workshop at the ASIS Conference. Washington, D.C.

Salton, G., C. Buckley and M. Smith (1990) "On the Application of Syntactic Methodologies in Automatic Text Analysis". *Information Processing & Management*, Vol. 26, No. 1.

Schank, R. (1976) *Conceptual Information Processing*. North Holland: Amsterdam.

Scholten, W. (1993) *'Library of the Future' takes shape at Columbia University Law Library.* Press release on the Empiricists Electronic Bulletin Board, CSLI.Stanford.EDU, Stanford, CA. Feb. 2, 1993.

Sedelow, S. and W. Sedelow (1986) "Thesaural knowledge representation," In *Proceedings, 2nd Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary: Advances in Lexicology,* University of Waterloo: Waterloo, Ontario.

Sells, P. (1985) *Lectures on Contemporary Syntactic Theories.* University of Chicago Press, Chicago.

Sharp, R.M. (1985) *A Model of Grammar Based on Principles of Government and Binding,* Master of Science Thesis, Department of Computer Science, University of British Columbia.

Smeaton, A. F. (1986) "Incorporating Syntactic Information into a Document Retrieval Strategy: an Investigation." In Fausto Rabitti, (Ed.), *Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval*, Pisa, Italy, September 8-10, 1986. Association of Computing Machinery, New York: 103-113.

Smeaton, A. F. (1989) *Using Parsing of Natural Language as part of Document Retrieval.* PhD Dissertation, University College Dublin.

Smeaton, A. F. and C. J. van Rijsbergen (1988) "Experiments on Incorporating Syntactic Processing of User Queries into a Document Retrieval Strategy." *Proceedings of the 1988 ACM SIGIR Conference on Research and Development in Information Retrieval.* Association of Computing Machinery, New York: 31-50.

Smeaton, A. F., A. Voutilainen and P. Sheridan (1990) *The Application of Morpho-Syntactic Language Processing to Effective Text Retrieval*, SIMPR-DCU-1990-165e, The SIMPR Consortium, ESPRIT Project 2083.

Somers, H.L. (1987) *Valency and Case in Computational Linguistics.* Edinburgh: Edinburgh University Press.

Sowa, J.F. (1984) *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley: Reading, MA.

Sparck Jones, K. and M. Kay (1973) *Linguistics and Information Science.* Academic Press, New York.

Sparck Jones, K. (1974) "Automatic Indexing", *Journal of Documentation,* 30(4): 393-432.

Saint-Dizier, P. (1989) "Programming in Logic with Constraints for Natural Language Processing," In *Proceedings from the European Association of Computational Linguistics*, 87-94.

Talburt, J.R. and D.M. Mooney (1989) "Determination of Strongly-Connected Components in Abstract Thesauri by the Method of Quartets," In *Proceedings of the Workshop in Applied Computing '89,* Oklahoma State University: Stillwater, OK.

Trier, J. (1932) *Sprachliche Felder.* Zeitschrift für deutsche Bildung. 8.417-27.

van Riemsdijk, H. and E. Williams (1986) *Introduction to the Theory of Grammar.* MIT Press: Cambridge, MA.

Vijay-Shanker, K. and A.K. Joshi (1985) "Some computational properties of tree adjoining grammars." In *Proceedings, 23rd Meeting of the Association for Computational Linguistics*, Chicago, July 1985.

Wang, Y-C, Vandendorpe, J. and Evens, M. (1985) "Relational Thesauri in Information Retrieval." JASIS 36 (1), 15-27.

Wehrli, E. (1990) "STS: An Experimental Sentence Translation System," In *Proceedings of COLING-90*, (1) 76-78.

Wilkins, E. (1988) *Syntax and Semantics: Thematic Relations,* Academic Press, Inc.: San Diego, CA.

Wittgenstein, L. (1934) *Philosophical Grammar.* trans. A. Kenney, University of California Press, Berkeley, CA. 1978.

Zubizaretta, M.L. (1987) *Levels of Representation in the Lexicon and in the Syntax,* Foris: Dordrecht.

☆U.S. GOVERNMENT PRINTING OFFICE: 1996-710-126-20173

MISSION

OF

ROME LABORATORY

Mission. The mission of Rome Laboratory is to advance the science and technologies of command, control, communications and intelligence and to transition them into systems to meet customer needs. To achieve this, Rome Lab:

a. Conducts vigorous research, development and test programs in all applicable technologies;

b. Transitions technology to current and future systems to improve operational capability, readiness, and supportability;

c. Provides a full range of technical support to Air Force Materiel Command product centers and other Air Force organizations;

d. Promotes transfer of technology to the private sector;

e. Maintains leading edge technological expertise in the areas of surveillance, communications, command and control, intelligence, reliability science, electro-magnetic technology, photonics, signal processing, and computational science.

The thrust areas of technical competence include: Surveillance, Communications, Command and Control, Intelligence, Signal Processing, Computer Science and Technology, Electromagnetic Technology, Photonics and Reliability Sciences.